Signal analysis and perceptual tests of vowel responses with an interactive source filter model

Nord, L. and Ananthapadmanabha, T. V. and Fant, G.

journal: STL-QPSR
volume: 25
number: 2-3
year: 1984
pages: 025-052

http://www.speech.kth.se/qpsr
B. SIGNAL ANALYSIS AND PERCEPTUAL TESTS OF VOWEL RESPONSES WITH AN INTERACTIVE SOURCE FILTER MODEL
L. Nord, T.V. Ananthapadmanabha, and G. Fant

Abstract
An interactive source filter model has been used to generate impulse responses, single formant sounds, and vowels. With the model it is possible to change physiological source parameters, such as lung pressure, peak glottal area, and glottal area function. The acoustic load is represented by the input impedance network of vocal tract and subglottal system. The four major acoustic effects of interaction: skewing, truncation, dispersion, and superposition are discussed and illustrated. A comparison has been made for an inventory of sounds obtained with and without the source-filter interaction. In a psychoacoustic discrimination test it was found that listeners, in a controlled situation, were able to perceive the fine temporal details for impulse and periodic single resonant sounds simulated with and without a vocal tract loading. However, for more complex sounds, such as male and female vowels, the discrimination approached chance level. Consequences of using an interactive model for the synthesis of connected speech are discussed.

Introduction
There has been a growing interest towards a better understanding of the voice source, especially the acoustic loading effect on the true glottal flow (Flanagan & Landgraf, 1968; Ishizaka & Flanagan, 1975; Guerin et al., 1976; Rothenberg, 1980; Fant, 1979; 1981; 1982; Ananthapadmanabha & Fant, 1982). However, there is a paucity of data on the perceptual consequences and the details concerning the temporal and spectral consequences of the interaction phenomena on the acoustic signal. In 1982 we reported a preliminary study on these aspects. See also a study by Childers et al. (1983).

A refined modelling of the voice source and a better theoretical understanding of the production of voiced speech sounds would hopefully lead to an improved quality of synthetic speech, especially of female and children voices. The aim of this work is a) to study the perceptual effects of interaction and b) to illustrate the temporal and spectral consequences.

Acoustic Theory
In the classical acoustic theory of speech production (Fant, 1960; Flanagan, 1965), it is assumed that the glottal impedance is very large compared to the vocal tract input impedance. Hence, the volume velocity air flow through the glottis, to a first approximation, is independent of the vocal tract loading. Any finite effect of the coupling is accounted for by changes in the bandwidth values of the vocal tract formant frequencies. In most of the earlier studies the effect of vocal tract filter loading on the aerodynamic glottal air flow source was underestimated. The time varying glottal impedance was approximated by quasi-static states, and further, a) the average dynamic resistance of the glottis was either compared with the resistance value of the vocal
tract filter at the formant frequencies, or, b) the impedance of the
glottis was compared with the vocal tract load at the pitch frequency.

A recent analytical approach by Ananthapadmanabha & Fant (1982) has
shown that the vocal tract equivalent network has to be analysed on a
transient theory basis by considering the time-varying glottal impe-
dance, see Fig. 1. In such a case, the appropriate impedances to com-
pare, when estimating the influence of the load on the glottal flow, are
the hypothetical glottal inductance and the input inductance of the
vocal tract load at the formant frequencies. However, for predicting
the formant damping due to glottal coupling, factor (a) above is still
valid.

On the basis of transient theory analysis, the following major
effects of source-filter interaction have been identified (Ananthapad-
manabha & Fant, 1982; Fant & Ananthapadmanabha, 1982). These are (i)
skewing, (ii) truncation, (iii) dispersion, (iv) superposition - linear,
(v) superposition - nonlinear, (vi) superposition - mechanical, and
(vii) supraglottal. The first four effects are based on the equivalent
circuit of Fig. 1. Here, it is assumed that the time-varying glottal
area function \( A_g(t) \) is given as input data (Flanagan, 1968), thus bypas-
sing the mechanical modelling of vocal fold vibrations. The input im-
pedances are represented by a Foster reactance network (Guerin et al.,
1976).

(i) Skewing: Under no load condition (short circuited load) the glottal
flow \( U_{sg}(t) \) is a scale factor times the glottal area function \( A_g(t) \). Due
to the load on the glottis, the flow pulse \( U_{sg}(t) \) is skewed compared to
the area function \( A_g(t) \), see Fig. 2. This skewing depends on the input
inductance of the load. Acoustically, the main effect of skewing is to
uniformly increase the level of all the formants. This can be inferred
from the derivative of the glottal flow pulse. (However, with extreme
supraglottal narrowing, as for stops, the pulse shape is smoothed out
due to the reduced transglottal pressure, and glottal abduction, sub-
sequently reducing the formant amplitudes.)

(ii) Truncation and (iii) Dispersion: The vocal tract resonances and
bandwidths undergo continuous modulation over the glottal open phase due
to coupling to the subglottal system and the time-varying nonlinear
 glottal impedance. The instantaneous bandwidth for F1 over a glottal
cycle is shown schematically in Fig. 3a for two vowels. As can be seen,
the bandwidth varies considerably, resulting in an exponential ringing
over the glottal closed phase and an almost complete truncation over the
open phase due to the increased glottal damping as is illustrated in
Fig. 3b. Due to the almost abrupt damping of the response waveform
during the open phase, the truncation effect can be described as the
weighting of the signal with a rectangular time window. The frequency
domain response will possess a \( \sin(x)/x \) shape, centered at the formant
frequency with clearly identifiable side lobe peaks. A comparison be-
tween a model-generated interactive (truncated) response and a non-
interactive (exponentially damped) response is made in Fig. 4. Two
examples of abrupt damping taken from selectively inverse filtered
samples of natural vowel sounds are shown in Fig. 5.

As the glottis opens, the resonances of the system undergo frequen-
cy modulation. This modulation is more due to an equivalent inductance
associated with the time-varying glottal conductance than to the physi-
ical inductance of the glottis (see Ananthapadmanabha & Fant, 1982). When
Fig. 1. Equivalent circuit over the open phase. (From Ananthapadmanabha & Fant, 1982, Fig. 4.)

Fig. 2. A source pulse and its derivative.

a) no load condition $U_{sc}(t)$,

b) with vocal tract load $U_g(t)$. 

---

Lung pressure

subglottal input impedance

time-varying glottal impedance

vocal tract input impedance

cc/sec

800

400

3 6 msec

DERIVATIVE

a b

3 6
Fig. 3. a) Typical instantaneous first formant bandwidth over a glottal cycle for two vowels.
   b) Interactive response of first formant circuit.
Fig. 4. Comparison of interactive (truncated) and non-interactive (exponentially damped) responses.

Fig. 5. Selective inverse filtering on natural speech samples. 
  a) Speech wave, b) selective inverse filtering. Top: 
    \( F_1 \sim 711 \text{ Hz} \), bottom: \( F_2 \sim 1226 \text{ Hz} \), c) estimated glottal flow.
the truncation of the response is smooth, the amplitude of the side lobes in the spectrum will depend more on the dispersion effect.

(iv) Superposition - linear. In those cases (notably high pitched voices) where the truncation is not significant, energy will be carried over from one glottal period to the next. Even for a perfectly periodic phonation, the vowel responses in the time domain for adjacent periods may thus not be identical. The degree of the superposition depends on the relation between pitch frequency and formant frequency.

(v) Superposition - nonlinear. Also the superposition within the open phase may affect the glottal flow derivative at the instant of closing discontinuity and, thus, a change in the excitation level (Fant & Ananthapadmanabha, 1982).

(vi) Superposition - mechanical: It is suspected that the presence of a superposition component may, at times, affect the mechanical vibrations of the vocal folds. Thus, a superposition component of the \( F_1 \) response may control the level of other formants as well (Fant et al., 1963; Fant & Ananthapadmanabha, 1982). However, it is not yet settled whether this is a nonlinear - acoustic or a mechanical interaction. Up till now simulations have not provided a support of the nonlinear component alone.

(vii) Supraglottal. It is evident that a supraglottal constriction will affect the transglottal pressure drop and, thus, the pattern of vocal cord vibrations. This is typical of voiced fricatives.

**Generation of Vowel Responses**

In this study we have used an interactive source-filter model with physiological parameters, such as lung pressure, glottal area function (maximal glottal area, open quotient*, area function), and pitch as input source parameters and the vocal tract input impedance as system parameter. See Fig. 6a for a schematic block diagram of the model. The mechanical modelling of phonation is not represented. Hence, by the above model the first four effects can be simulated. With the model it is thus possible to independently change the input parameters and study the effects of, e.g., a raised subglottal pressure.

**EXPERIMENTS**

In this study we have taken into account the coupling effect of \( F_1 \) load only with no subglottal coupling.

Conventional formant synthesis of vowel sounds uses constant formant bandwidths resulting in an exponential damping, as already shown in Fig. 4, with more pronounced superposition effects than found in natural speech.

---

* open quotient = the ratio of the glottal open time to the fundamental period.
Fig. 6. Block diagram of source-filter models.
Here, we have focused on the ability of the auditory system to discriminate between the two types of responses, viz., the exponential impulse response with fixed resonant frequency and bandwidth and the response of a resonance system with truncation and dispersion effects. We shall refer to the nonlinear response as the "interactive response" of the formant because of source filter interaction.

Bandwidth values have been chosen for the exponentially damped stimuli so as to obtain about the same loudness level for the two types of signals. Perceptual experiments by Fant & Liljencrantz (1979) have shown that listeners match to an "effective bandwidth", resulting in equal mean values of the two temporal envelopes (exponential and gated). This will ensure the spectral level at the formant peaks to be the same. On the concept of bandwidths, see Fant (1982) and Fig. 7.

TEST PROCEDURE
Perceptual tests were conducted using an ABX paradigm, where A, B, or X could be the "interactive response" or the "exponentially damped response". A test list consisted of randomly ordered ABX stimuli. The same ABX triplet was repeated once giving the listener a choice to indicate the level of confidence or to reverse his judgement. Time gaps of 250 msec separated A, B, and X, with a time interval of 4 sec. to the repetition for confidence level judgement. The new ABX triplet was presented after 7 sec. The listeners heard stimuli with the same first formant frequency throughout a given list. The stimuli were presented over headphones at a comfortable listening level. One test list lasted approximately 5 min. A maximum of five lists were presented during one listening session. Experienced listeners participated in the test.

SPECIFICATION OF STIMULI
We used three types of signals: impulse responses of a single formant circuit, periodic single formant sounds, and vowels. Thus, we could study the discriminability as a function of the complexity of the signal.

Formant frequencies and bandwidths for closed glottis conditions and vocal tract inductance values were taken from Ananthapadmanabha & Fant (1982). These values were combined with representative values of glottal source parameters, see Table I and Fig. 8. As only first formant loading was considered for the synthesis of vowels, the effective bandwidths for B₂ to B₅ were used.

* In a preliminary report (Ananthapadmanabha et al., 1982) we used the term "true response" referring to the resemblance to formant oscillations found in natural speech.
Fig. 7. Spectral effect of truncating a formant oscillation: Bandwidth definitions (from Fant, 1982, Fig. 10 modified). Log spectra of a) exponential response with closed phase bandwidth ($B_c$), b) truncated response with 3 dB down points ($B_t$), c) exponential response with effective bandwidth ($B_{eff}$).

Fig. 8. Specifications for the single resonant stimuli. A natural grouping is indicated.
TABLE I.
Stimuli specifications for single resonant responses and male vowels.
(Formant frequencies and bandwidths in Hz, inductances in mH)

<table>
<thead>
<tr>
<th></th>
<th>F_1</th>
<th>F_2</th>
<th>F_3</th>
<th>F_4</th>
<th>F_5</th>
<th>B_1</th>
<th>B_2</th>
<th>B_3</th>
<th>B_4</th>
<th>B_5</th>
<th>L_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>659</td>
<td>1060</td>
<td>2418</td>
<td>3490</td>
<td>4087</td>
<td>32</td>
<td>31</td>
<td>56</td>
<td>152</td>
<td>54</td>
<td>6.14</td>
</tr>
<tr>
<td>/o/</td>
<td>524</td>
<td>846</td>
<td>2357</td>
<td>3428</td>
<td>3969</td>
<td>35</td>
<td>24</td>
<td>28</td>
<td>32</td>
<td>48</td>
<td>5.73</td>
</tr>
<tr>
<td>/u/</td>
<td>277</td>
<td>610</td>
<td>2373</td>
<td>3683</td>
<td>4022</td>
<td>70</td>
<td>23</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>7.97</td>
</tr>
<tr>
<td>/i/</td>
<td>269</td>
<td>2257</td>
<td>2876</td>
<td>3648</td>
<td>4260</td>
<td>63</td>
<td>22</td>
<td>26</td>
<td>80</td>
<td>43</td>
<td>7.31</td>
</tr>
<tr>
<td>/e/</td>
<td>444</td>
<td>1940</td>
<td>2716</td>
<td>3356</td>
<td>4092</td>
<td>31</td>
<td>49</td>
<td>192</td>
<td>200</td>
<td>196</td>
<td>2.77</td>
</tr>
</tbody>
</table>

not used

TESTS ON IMPULSE RESPONSES

Impulse response sounds were generated using the interactive model with only one resonance. Five sets of F_1, B_1, L_1 values were taken from Table I. A constant subglottal pressure of 8 cm H_2O and an open quotient value of 50% were used. The specification of the open quotient in this case means that although only one instant of excitation occurred in the simulation, the model calculated the interaction effects for as long a time as there was any energy left of the formant ringing. Thus, the resulting wave could in some cases contain a second interval of (low energy) ringing stemming from the next closed glottis interval.

For each stimulus the effective bandwidth was calculated and used for the exponentially damped response so as to ensure the same loudness level of the stimuli.

In order to simulate different degrees of glottal damping, three values of peak glottal area, viz., 30, 20, and 10 sq.mm were used. As the peak glottal area decreases, the damping over the open phase decreases and the envelope contour of the interactive response approaches that of the exponential response. In Fig. 9 temporal responses and spectral sections of parts of the stimuli sets are compared. The sin(x)/x shape with side lobes is clearly seen for the interactive response. The levels of the side lobes decrease as the glottal area diminishes.

The results of the listening tests are shown in Table II. This Table gives the number of errors for simulation with each individual peak glottal area as well as the pooled percentage correct discrimination. The discrimination is well above chance level for all the stimuli. As the peak glottal area decreases, the discriminability also decreases,
Fig. 9. Comparison of interactive and non-interactive responses in time and frequency domains: Impulse response stimuli.

a) $f_1$ of /a/, simulated with three values of peak glottal area.
Fig. 9. Comparison of interactive and non-interactive responses in time and frequency domains: Impulse response stimuli.

b) Stimulus with low $F_1$ ($F_1$ of /i/)
c) Stimulus that gave the lowest discrimination score ($F_1$ of /e/).
as expected. But, for the stimuli with high $F_1$, the discriminability is high, even for a low peak glottal area of 10 sq.mm.

**TABLE II**

Result of discrimination test: Impulse response of a single resonance

Number of errors and percentage correct discrimination. Three values of peak glottal areas.
(20 samples/data point, 5 listeners)

<table>
<thead>
<tr>
<th>Formant frequency of the resonance (Hz)</th>
<th>Number of errors</th>
<th>Correct response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak gl. area (sq.mm.) (pooled)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>92</th>
</tr>
</thead>
<tbody>
<tr>
<td>659</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>95</td>
</tr>
<tr>
<td>524</td>
<td>0</td>
<td>10</td>
<td>7</td>
<td>72</td>
</tr>
<tr>
<td>444</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>85</td>
</tr>
<tr>
<td>277</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>80</td>
</tr>
<tr>
<td>Sum:</td>
<td>4</td>
<td>19</td>
<td>23</td>
<td>mean: 85</td>
</tr>
</tbody>
</table>

The exponential stimuli sounded somewhat resonant in quality due to longer ringing, while the interactive response stimuli sounded numb, like tapping on a non-resonant piece of wood.

Theoretically, the relative amount of truncation effect (RTE) depends on the values of the parallel resistance $R_1$ of the vocal tract load at $F_1$ and the peak glottal resistance. This can be expressed with the following formula:

$$\text{Relative truncation effect (RTE)} \sim (F_1^2 \cdot L_1/B_1) \cdot (A_{\text{max}} \cdot T_{\text{open}}) \quad (1)$$

where

$F_1, B_1$ = formant frequency and bandwidth, respectively

$L_1$ = inductance load of the first formant

$A_{\text{max}}$ = peak glottal area

$T_{\text{open}}$ = duration of the glottis open interval

$R_1 \sim (F_1^2 \cdot L_1/B_1)$ for the single resonance load.
A large RTE-value would mean that the difference between the interactive and the non-interactive stimulus is enhanced. In Fig. 10 the number of errors is plotted as a function of the RTE-value. The tendency of getting better discrimination scores for large RTE-values is clearly seen.

Some characteristics of the interactive response stimuli should be noted like the shape of the broad peak and the pronounced side lobes. For some stimuli the side lobes are only 15 dB weaker than the main peak. Moreover, it is evident from the discrimination results that listeners are able to perceive very small differences between stimuli pairs.

TESTS ON SINGLE FORMANT SOUNDS

Periodic single formant sounds of 400 msec duration with a smooth intensity onset and offset were synthesized. Conventionally, a sequence of impulses constitutes the driving source in the synthesis of vowel sounds. But, here we used source pulses specifically computed with the interactive source-filter model.

Natural vowels and, thus, also single formants are more complex than impulse responses. This is due to a complex of components (Fant, 1979): (i) Glottal flow pulses in a differentiated form appear in the waveform, (ii) there may be an additional excitation at the instant of glottal opening, (iii) due to the superposition effect tail ends of the damped waves from previous periods may add in arbitrary phase to the main response of the present period. The presence of these additional components could reduce the perceptual discriminability results.

A peak glottal area of 20 sq.mm and a constant subglottal pressure of 8 cm H2O were used as source parameters. Two values of pitch, \( F_0 = F_1/N \) and \( F_0 = F_1/(N-0.5) \), N-integer, were used for each sound. These two cases correspond to a pitch harmonic coinciding with the formant peak or to the harmonics lying on either side. Three different values of open quotient: 30 (or 40), 50, and 70% were used, giving six different stimuli specifications for each sound. The effect of subglottal load was neglected.

Figs. 11a and 11b display the single formant stimuli in the time and frequency domains. For vowels with high \( F_1 \), the differentiated glottal pulse gives rise to a spectral peak, marked \( F_G \). As the open quotient value is increased, the source spectral peak is lowered in frequency. For stimuli with a low \( F_1 \) the source spectral maximum and formant peak merge and are difficult to distinguish. This can be seen both in the time and in the frequency domain. The side lobes for the single formant sounds are not as apparent as for the impulse stimuli, compare Fig. 9.

Results of the listening tests are shown in Table III. The results once again show a good discriminability, especially for sounds with a high first formant. It may be noted that for stimuli with \( F_0 = F_1/(N-0.5) \), the discriminability is relatively better, except for stimuli with \( F_1 = 269 \) Hz. This could be explained as follows: When \( F_0 = F_1/N \) the
Fig. 10. Number of errors for the discrimination of impulse responses as a function of the relative truncation effect. 20 samples/data point.
Fig. 11. Comparison of interactive and non-interactive responses in time and frequency domains: Periodic single formant stimuli. $F_0 = F_1/N$. A phonetic judgement is indicated.

a) $F_1$ of /a/, simulated with three degrees of open quotient.
Fig. 11. Comparison of interactive and non-interactive responses in time and frequency domains: Periodic single formant stimuli. $F_0 = F_1/N$. A phonetic judgement is indicated.

b) Stimulus with low $F_1$ ($F_1$ of /i/), simulated with three degrees of open quotient.
superposition effects overrides the effect of truncation thus reducing the differences between the stimuli pairs. As there was no correlation between open quotient values and the number of discrimination errors, the results are pooled over the open quotients.

**TABLE III**
Result of discrimination test: single formant sounds

Number of errors and percentage correct discrimination, pooled for two pitch conditions.
(72 samples/data point, 6 listeners.)

<table>
<thead>
<tr>
<th>Formant frequency of the resonance (Hz)</th>
<th>Number of errors</th>
<th>Correct response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_0 = F_1 / N$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_0 = F_1 / (N-0.5)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>659 (N=7)</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>444 (N=4)</td>
<td>36</td>
<td>17</td>
</tr>
<tr>
<td>269 (N=3)</td>
<td>22</td>
<td>26</td>
</tr>
</tbody>
</table>

Sum: 68 46 mean: 74

Open quotient = 30 (or 40), 50 and 70 %, $P_{subgl.} = 8$ cm H$_2$O

**PHONETIC QUALITY**

If the spectral shape of single resonant sounds are vowel-like with a realistic pitch, they will have a (weak) phonetic quality, that is, listeners tend to perceive the sounds as phonemes. The non-interactive response for stimuli with $F_1 =$ 659 and 269 Hz (Table III) were, thus, identified as the vowels [ae] and [u], respectively. The side lobes appearing for the interactive response stimuli with the same $F_1$ and a high open quotient value were enough to shift the quality towards the phonemes [a] and [o], respectively. Referring to the spectral sections of the stimuli, see Fig. 11a (lower) and 11b (lower), this effect can be interpreted if we assume the side lobe peaks as contributing to a shift of gravity in the spectra.
**TESTS ON MALE VOWELS**

Low pitch male vowels were synthesized using five formants and an intonation pattern modelled after a natural speech sample. Sampling frequency = 16 kHz. A natural intensity contour was obtained (in a somewhat artificial way) by varying the peak glottal area (0-20 sq.mm) keeping the subglottal pressure constant (P_s = 8 cm H_2O, Duty Ratio = 50%). No subglottal coupling was used. For vowels with an exponentially damped first formant, a constant effective bandwidth giving a good average fit was chosen. Spectral sections of the stimuli are shown in Fig. 12. The voice source maximum is seen for vowel /a/ and /o/, but is merged with F_1 for /u,i/ and /e/. Spectral differences between the interactive and the exponential responses are small. The test results (see Table IV) also give low correct discrimination scores, and for some vowels just above chance level. (In a preliminary study by Ananthapadmanabha et al., 1982, we reported somewhat better discrimination scores. This was due to a bad choice of bandwidths for the higher formants which resulted in some artefacts in the stimuli.)

Listeners found the task to be very difficult, although they reported that the stimuli sounded natural, the only cue being "small voice quality changes".

**TABLE IV**

Results of discrimination test: male vowels

Number of errors and percentage correct discrimination.
(96 samples/data point, 6 listeners)

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Number of errors</th>
<th>correct response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>42</td>
<td>56</td>
</tr>
<tr>
<td>/o/</td>
<td>44</td>
<td>54</td>
</tr>
<tr>
<td>/u/</td>
<td>31</td>
<td>68</td>
</tr>
<tr>
<td>/i/</td>
<td>41</td>
<td>57</td>
</tr>
<tr>
<td>/e/</td>
<td>37</td>
<td>61</td>
</tr>
</tbody>
</table>

mean: 59
PERCEPTUAL EFFECTS OF \((F_1 + F_2)\)LOAD

We also simulated a number of stimuli taking into account the interaction of both the first and the second formant loads, in order to see whether the perceptual effects would be stronger. This turned out not to be the case. We conclude that the main results in this study will apply equally well for a modelling with the loading of the complete vocal tract network.

TESTS ON FEMALE VOWELS

The same vowels were synthesized with a high pitch and using average formant frequency values for female speakers, taken from Fant (1972), see Table V. The input inductances were assumed to be the same as in Table I for the male data, since there is a decrease of the length as well as the cross-sectional area of the vocal tract. In order to obtain a female voice quality, we found it necessary to use higher subglottal pressure, smaller peak glottal area \((0 - 14 \text{ sq.mm})\), and a higher open quotient value.

TABLE V

Stimuli specifications
(Formant frequencies and bandwidths in Hz, inductances in mH)

\[
\begin{array}{cccccccc}
F_1 & F_2 & F_3 & F_4 & B_1 & B_2 & B_3 & B_4 & L_1 \\
/a/ & 700 & 1100 & 3000 & 4200 & 60 & 66 & 114 & 150 & 6.14 \\
/u/ & 365 & 690 & 2700 & 4200 & 96 & 60 & 70 & 120 & 7.97 \\
i/* & 345 & 2200 & 3100 & 4200 & 80 & 70 & 105 & 120 & 7.31 \\
/e/ & 395 & 2200 & 2850 & 4200 & 50 & 67 & 190 & 200 & 2.77 \\
\end{array}
\]

Values taken from Fant (1972) except *).

\(F_5 = 5.4 \text{ kHz, } F_6 = 6.6 \text{ kHz, } F_7 = 7.8 \text{ kHz}\)

\(B_5 = 300 \text{ Hz, } B_6 = 400 \text{ Hz, } B_7 = 500 \text{ Hz}\)

Sampling frequency= 16 kHz

Open quotient= 60 \%, \(P_{\text{subgl}} = 10 \text{ cm H}_2\text{O}, \)

\(F_{\text{max}} = 265 \text{ Hz}\).
The female voice quality was good, although we did not optimize the parameter settings. Vowel spectra are shown in Fig. 13. For the vowel /ɛ/ a glottal zero is seen at about 800 Hz, both in the interactive and non-interactive simulation. This zero originates from the combination of two excitations with the particular choice of open quotient. For this simulation, the second excitation at the instant of glottal opening turned out to be quite strong. A change of open quotient by a few percent removed the glottal zero completely.

The result of the listening test, see Table VI, shows approximately the same discrimination rates as for the male vowels, that is, a little above chance level.

**TABLE VI**
Result of discrimination test: female vowels

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Number of errors</th>
<th>correct response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>34</td>
<td>65</td>
</tr>
<tr>
<td>/u/</td>
<td>38</td>
<td>60</td>
</tr>
<tr>
<td>/i/</td>
<td>22</td>
<td>77</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>41</td>
<td>57</td>
</tr>
</tbody>
</table>

mean: 65

**Discussion**

A comparison of the discrimination test results for the three types of signals is shown in Fig. 14. Though the discrimination was easy for signals with simple structure, the task became difficult for vowel sounds. The result is not surprising. For vowel sounds, these interaction effects could be considered to be of lesser importance compared to the dynamic variation of the glottal pulse shape which contributes more to the voice quality (Rosenberg, 1970; Fant, 1979; Ananthapadmanabha 1984).

However, this should not be misinterpreted as saying that source-filter interaction is of no consequence for the synthesis of connected speech. In the above psychoacoustic experiments we have tested the discriminability for fine temporal differences, controlling all other parameters to be the same. Thus, we have used carefully calculated effective bandwidth values for the exponentially damped cases to keep the loudness level the same between stimuli pairs.
Fig. 13. Spectral sections of female vowel stimuli.
Fig. 14. A comparison of discrimination results for the different stimulus types.

Stimuli Type

<table>
<thead>
<tr>
<th>Format</th>
<th>Vowels</th>
<th>Vowels</th>
<th>Single Vowels</th>
<th>Tones</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clicks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correlation Level

0

Chance Level

50

100
For natural speech the interaction effect will result in a simultaneous change of the spectral shape (i.e., $\sin x/x$) and the formant levels. The above psychoacoustic experiment has shown that if the formant levels are correctly controlled then the effect of spectral shape differences is small.

For the synthesis of connected speech (other than steady state vowels), the loudness contrasts between speech segments can be maintained only by using carefully calculated effective bandwidth values. The effective bandwidth depends not only on the articulation but also on the source parameters. Values as large as 400 Hz for $F_1$ of a voiced /h/ sound are not uncommon. Similarly, $B_2$ of /i/ could be influenced strongly by the truncation effect. Because of these interaction phenomena, it becomes difficult to predict the effective bandwidths by rule. The existing rules for deriving bandwidths based on the formant frequencies are for the closed glottis condition (Fant, 1972). Furthermore, in practical synthesis of high pitched voices, the bandwidth sometimes has to be modified (increased) in order to avoid unrealistic amplitude fluctuations due to superposition effects.

This gives us two options of synthesis strategy. Either a) The effective bandwidth has to be carefully determined and used in a non-interactive model or b) available rules for closed phase bandwidths can be used in an interactive model. The later alternative is preferable as it models the speech production process to a better accuracy, including a proper control of the superposition effect, which is especially critical for the synthesis of high pitched voices. A third alternative (possibly of technical interest) would be to modulate the bandwidth within each glottal cycle, according to source-filter interaction theory, but still use a non-interactive synthesis model.

SOME RELATED TOPICS

The above experiments have led us to inquire into some related topics which we will briefly discuss.

A. Perception of $F_1$

What is the perceptual significance of a fusion of spectral voice source maximum and a low first formant peak in real speech? Is the perceived first formant location affected by a strong source peak as seen in the spectral sections, or does our hearing neglect the contribution from the source peak? There is some evidence for the influence of the spectral shape on the perceived first formant frequency (Chistovich & Ogorodnikova, 1982). But there is also some counter evidence based on an auditory short term spectral analysis (Schroeder & Mehrgardt, 1982; Millar & Underwood, 1975).

B. Short term spectral analysis

When discussing the spectral components of signals that exhibit a fine temporal structure, a harmonic (long term) spectrum will not clear-
Fig. 15. Consecutive short term spectra for the interactive response of a single formant sound. (T = 10 msec, Hamming window, 1 msec displacement between sections.)
ly reveal the variations. Another way of representation is to use a running short term spectral analysis. In Fig. 15 a number of consecutive short term spectra for a single formant sound are shown, covering approximately one and a half glottal cycle. These spectra were obtained by computing the FFT on Hamming windowed segments of 10 msec duration with 1 msec steps between the sections. The main features to be observed are the changing level of the voice source maximum and the \((\text{sin}x/x)\) shaped contour of the \(F_1\) peak with side lobes. With this type of analysis, the first formant peak can be expected to appear even for high pitched voices, viz., at those instances when the short term window includes the closed glottis interval.

**Conclusions**

The auditory system is known to be sensitive to differences in the waveforms, e.g., AM/QFM (Lozhkin, 1971). In vowel production, the coupling of the vocal tract to the time-varying glottal impedance gives rise to a non-exponential formant damping, compared to an exponential damping used in conventional vowel synthesis. We have tested the discriminability for non-exponential (interactive) and exponential (non-interactive) formant damping for a) impulse responses, b) periodic single formant sounds, and c) male and female vowel sounds. These two types of synthesized signals were easily discriminable for simple types of stimuli, though difficult for vowels sounds, Fig. 14.

The quality of synthetic speech cannot be improved very significantly by using an interactive source-filter model. However, the interactive model has the advantage of a simpler control strategy for bandwidths and it has a better control of the superposition effect. A more rigorous evaluation of the interaction phenomena is under way.

**References**


