SEARCH FOR UNEXPLORED EFFECTS IN SPEECH PRODUCTION

C.H. Coker, M.H. Krane, B.Y Reis and R.A. Kubli

Bell Laboratories, Murray Hill, NJ 07974
(B.Y. Reis is also a student at M.I.T.)

ABSTRACT

Speech coders invariably spend about 1/3 of their bits replicating a number of effects collectively termed "excitation". The need for so much bandwidth stems from two causes. The frame rate must be relatively high, because transient changes must be resolved. The data needed for each frame is high because unpredictable broad-band components must be reproduced. Here we discuss three projects to learn more about these elusive aspects of speech. One project models the transient behavior. Two others seek to characterize stochastic processes that accompany periodic vibration in voiced sounds.

1. Sequential effects in excitation

Figure 1 (a) shows a waveform of the beginning of the word heed. Fig. 1 (b) is probable glottal behavior based on data for a similar utterance — word-initial /hə/. The instrumentation is photo-glotography (PGG), also known as transillumination. A fiber-optic light source is inserted through the nose, and positioned above the glottis. In a darkened room, the outline of the glottis can be seen on the lower larynx and pharynx. A light sensor "watching" about 5 cm^2 in this region gets a reasonably linear response with glottal area. Fig. 1 (c) is Linear Prediction residual for the waveform in (a). Height of the sharp impulses reflect abruptness of glottal closure.

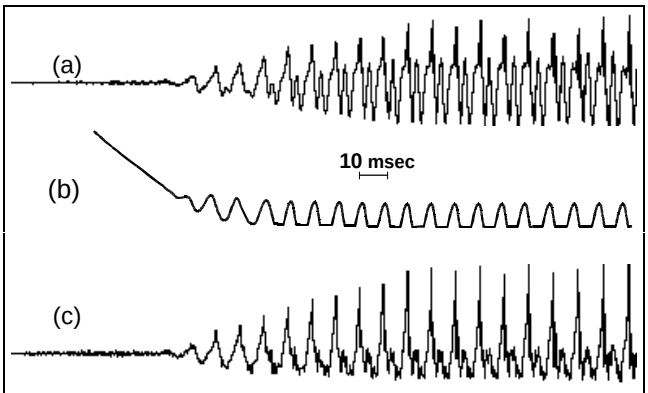

Figure 1 For the /hi/ in "heed"; (a) waveform; (b) approximate glottal (inferred from PGG for /hə/); (c) Linear Prediction residual.

In Fig. 1, the glottis is open for /h/; then closes for the vowel. Glottal vibration starts shortly before glottal "rest area" gets roughly to zero. In optical evidence, glottal area seems to reach a rather stable state perhaps 25 msec after vibration begins. However, the waveform and prediction residual show substantial change for another 60 msec. Something in the larynx is still changing. But what?

Here is offered a physiological explanation. A model is developed and its coefficients are tuned from available data. Although the model was suggested by a proposed explanation, it does not depend on details of that proposal. The model is anchored only in the notion that measurable glottal area and LP residual derive from a single variable, and conversely that behavior of that variable can be inferred from glottal instrumentation and LP data.

Fig. 2 is a sketch of the glottal mechanism. The L-shaped structures shaded gray are the *arytenoid cartilages.* The vocal cords attach to the tips of these cartilages, the *vocal processes*. The glottis is closed by tensing several muscles — principal among them, the *exterior thyro-arytenoids*. As tension increases, the arytenoids rotate and the vocal cords move together. At some point, however, the opposing arytenoid tips press together, and can move no further.

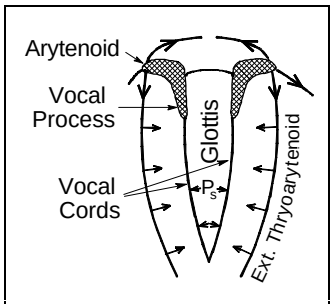

Figure 2 The glottal mechanism.

But there is a mechanism for the thyro-arytenoids to have effect after arytenoid motion is blocked. The cords are pushed apart in the middle by air pressure, and by stretch forces tending to restore the breathing state. The exterior thyro-arytenoids are curved and push inward through a draw-string action. Increasing tension, after the arytenoids block, pushes the cords in at the sides. This has a modest effect on glottal area. But a more profound effect is to alter the geometry of glottal closure.

Fig. 3 shows the progress of glottal closure for three adjustments. The lower left is a magnified view of glottal area vs. time for conditions a, b and c; the lower middle shows glottal area derivative, and lower right, the spectral consequences. Condition a is a partially open glottis as might occur in /z/ (exaggerated). The arytenoid tips are separated a small distance, making the vocal-cord configuration somewhat triangular. During vibration, vocal-cord closure begins at the front and progresses toward the arytenoids in the back. This "zipper" action has a time course of glottal area closely resembling a decaying exponential. This contributes a 6 dB/octave spectrum down-turn at a frequency F_T varying roughly inversely with measurable glottal area (see Fig. 5).

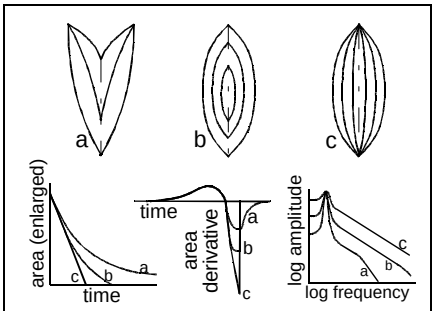

Figure 3. Modes of glottal closure.

Condition b is the case when the arytenoid tips just touch. Closure typically starts at both ends and progresses toward the middle. Final closure is still abrupt to a degree, depending on the distribution of side force along the cords. Condition c is the fully-closed adjustment for a stressed vowel. Closure occurs nearly simultaneously along the full length of the cords. Of course, the high-frequency sound is directly tied to the abruptness of glottal closure; progressions for a to b to c produce serious change in high-frequency amplitude.

The intention is to account for all of these phenomena with a single variable. There are two candidates for this variable, as well as some intermediates between the two. One choice would be to use the spring-force-neutral area A_{g0} as this primary variable. But A_{g0} saturates and changes little in the range where acoustic change is still high. And direct measurement of A_{g0} is thwarted by glottal vibration, specifically in that region where small mechanical changes are accompanied by large acoustic change.

The simpler option is to assume that A_{g0} is a nonlinear result of an underlying variable — call it glottal width control A_{gw}. We scale A_{gw} to agree with transillumination when there is no vibration (Fig. 4). Where there is vibration, we infer A_{gw} from prediction residual, suitably scaled and translated to make a smooth transition at the voiced/voiceless boundary.

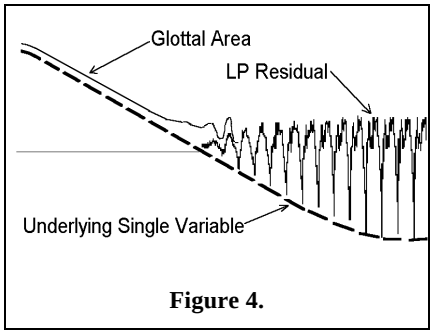

Figure 4.

With instrumentation for both open-glottis and voiced conditions, it's straight-forward to find how this variable behaves for sounds other than /h/. And it's easy enough to make rules that characterize the variable for all phonemes, both stressed and unstressed.

Fig. 5 (top), plots several acoustic quantities as functions of the glottal-width variable. Lines in the lower figure show the range of values of that variable for different phoneme groups. (Plus signs indicate values typical of stressed allophones.) For stressed vowels, tightly closed glottis makes for stronger high-frequency voicing. Stress-initial consonants have wider glottal opening than unstressed ones. This makes for higher intra-oral pressure and thus louder frication. In the CD version of this paper, graphic CD-1 [IMAGE A1011G01.GIF] shows a spectrogram, waveform and prediction residual for the sentence "Where there is a will there is a way." In stressed vowels of the words "where", "will" and "way", primary stress shows up as "spikiness" in the waveform and prediction residual, and strong vertical pitch striations in the spectrogram. Compare the stressed vowels with the reduced vowels of "is" and the two instances of "a". Very breathy voicing can be seen in the two examples of consonant /z/, and in the intervocalic semi-vowel consonants /r/ and /w/. Note that glottal transitions for /z/, /r/ /w/, etc. extend well into the adjacent vowels.

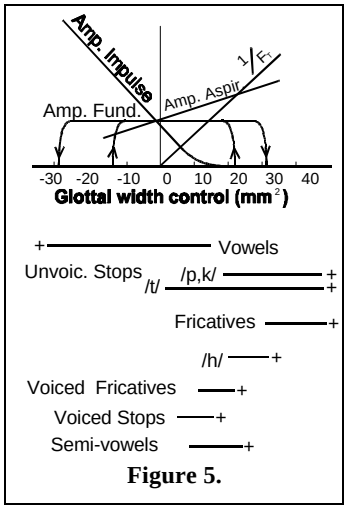

Figure 5.

Thus, a single slowly moving variable accounts for a wide range of transient, sequential and at-times prosodic and phonemic changes in several aspects of speech excitation.

2. Aspiration during voicing

Another project investigates a source of aperiodicity during voicing. Today it's common to excite synthesizers with a shaped pulse plus a burst of noise. Many believe this has a counterpart in real speech. And there is evidence from listening tests that noise overlapping the pulse is beneficial [Hermes, Speech Comm. 10 (1991) 496-502]. We can see partially non-repetitive components of LP residual, but measurement is, at best, approximate. To make accurate measurements, we turned to a mechanical model.

In the apparatus, a molded-rubber model glottis is mechanically driven. Under control of a phase-locked loop, it is closely synchronized with a pulse train from the computer. With pitch an exact submultiple of the sampling rate, many pitch periods can ensemble averaged to find the periodic component. That component is subtracted away, leaving only the noise. The square of that noise is again ensemble averaged, to obtain noise power vs. time.

Fig. 6 is a front cross-section view of the molded glottis. The vocal-tract and trachea fittings of the mold are interchangeable, so that direction of flow through the wedge-shaped glottal constriction can easily be switched. The difference between converging and diverging glottal geometry did not prove to be a first-order effect.

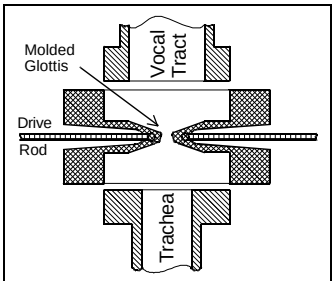

Figure 6 The glottal assembly.

For most experiments the vocal tract was represented by a straight tube 2.5 cm diameter by 17 cm long. The straight tube allows access for a hot-wire anemometer probe on an XYZ positioner, and access for a movable obstruction in the glottal jet. A trachea tube 1.5 cm dia. by 18 cm long opens abruptly into a 2.5 cm dia. hose. The hose, representing the lungs, is 10 meters long, with a tapered sound-absorbing wedge to form a non-reflecting sound termination. This configuration matches human data quite well. It produces closed-glottis resonances in the trachea at odd multiples of about 480 Hz. An area ratio if ~2.5 between lung and trachea matches the bandwidths. CD 2 [IMAGE A1011G02.GIF] shows waveforms of trans-glottal pressure. The upper trace is natural speech, measured by Cranen [Cranen & Boves, "Pressure Measurements During Speech Production…" JASA 77, 1543-1551 (1985)]. The lower trace was measured from the model.

Noise produced by the glottis alone is small compared to the periodic component. But an obstruction in the pulsating glottal jet makes the noise ~15 dB louder — comparable in level to an /h/. Several different shapes of obstructions were tested in both straight and bent tubes. The data shown here used a straight tube with a movable obstruction — an object shaped like a half thumb tack, positioned to block about half of the jet. Fig. 7 shows noise measured for a number of positions of the obstruction. The upper trace is glottal area as measured by transillumination.

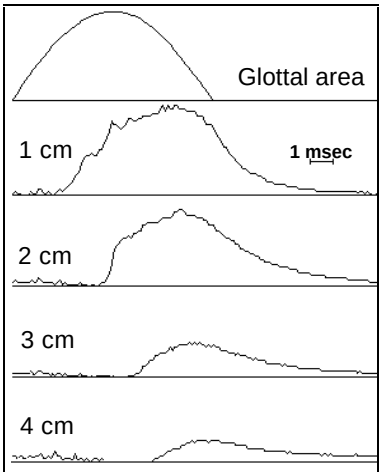

Figure 7. RMS noise vs. time for several distances from glottis to obstruction. (P_g = 7 mbar; pitch = 60 Hz.)

Noise changes in amplitude and phase with position of the obstruction. Fig. 8 shows peak noise vs. driving lung pressure, at several distances from the glottis. Noise power goes as about the 2.5 power of transglottal pressure; this corresponds to the 5th power of particle velocity ($p = \frac{1}{2}\rho v^2$).

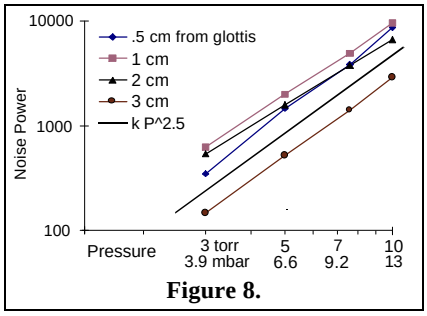

Figure 8.

Fig. 9 shows noise vs. distance from glottis to obstruction. Noise grows slightly with distance out to about 1.5 cm, the point where the entire burst has become turbulent. Then it drops at a rate of ~4 dB/cm.

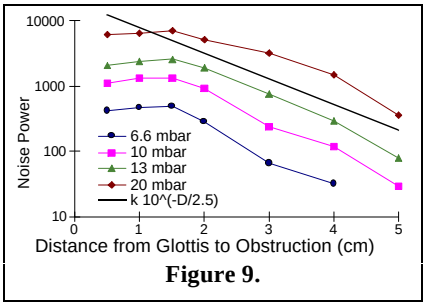

Figure 9.

In the noise-amplitude plots (Fig. 7), note that duration of the noise burst gets shorter with distance from the glottis. There are two causes. One is time-varying transglottal pressure [IMAGE A1011G02.GIF]. Pressure is low at the beginning of glottal opening and higher at closure, by a ratio of about 4:1. Air velocity goes as the square root of the driving pressure; air at the tail of the burst moves twice as fast as that at the front. Eventually it must catch up. There is also another effect.

Measurements with a hot-wire anemometer show velocity of 40 m/sec — about Mach 0.12 — at the trailing edge of the burst. Velocity at the leading edge is about half that: 20 m/sec. But in Fig. 7, the leading edge of the burst appears to advance at less than 10 m/sec (4 cm in about 5 msec). The glottal burst is a fast-moving packet of air crowding into slower air in front. As particles at the front must share momentum with the slower particles, they slow down, leaving particles further back to then be the leading edge of the velocity burst. Thus the apparent velocity in Fig. 7 is roughly an actual particle velocity of about 2 cm/msec minus about 1 cm/msec, the rate at which the front of the burst is being "eaten away" by loss of energy to slower air.

These experiments show that an obstruction ~2 cm from the glottis produces the most noise, and has delay about optimum to match LPC and Hermes' perception results. The likely candidate for this obstruction is the epiglottis. MRI and X-ray images often show the epiglottis positioned out in the middle of the trachea — not pushed against the tongue root as it would be in breathing [c.f. Narayanan, Alwan and Haker, "An articulatory study of fricative …", JASA 98 (3), Sept. '95]. This position is about optimum to block half the glottal jet — thus to produce the most aspiration during voicing.

3. Effects of air flow on acoustics of the vocal tract

The third project looks for fluid-dynamic effects inside the vocal tract. To serve as a check for computational experiments [Levinson, also this session], frequency responses of a vocal tract model were measured with and without DC air flow. The model (Fig. 10) was constructed of plastic cut and molded with filler, to form the back wall and tongue; then sandwiched between parallel plates. A shape for the vowel /u/ was chosen as one likely to produce fairly high particle velocities inside a complex shape.

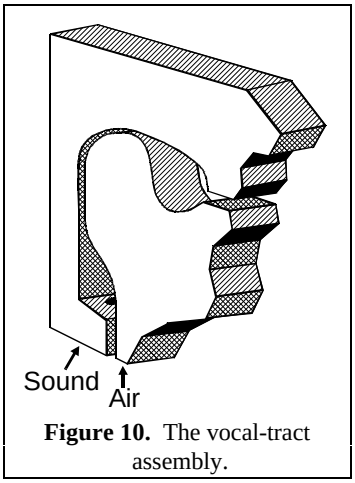

Figure 10. The vocal-tract assembly.

Air is injected through a non-reflecting tube to approximate the lungs. A probe tone is injected into this model at the point of the esophagus — about 2.5 cm downstream from the glottis. Again, the method provides the means to separate periodic from stochastic components. Analysis uses a Stanford Research SR770 FFT Network Analyzer. The device puts out a digitally repeatable sweep tone or burst of noise, then does a DFT of the return signal. The complex spectra are averaged over a specified number of repetitions. Sound not coherent with the probe tone gets averaged away (diminished by $1/\sqrt{N}$, the number of sweeps).

The probe-tone source is a heavy power loudspeaker driver, coupled through a 2.5 mm dia. tube, damped with a length of polyester yarn. A frequency response of the driver is captured in advance, and later used to divide the measured results — in effect de-convolving and removing the driver response.

Spectra were taken for the vocal tract with air flows between 0 and 860 ml/sec. The highest rate is roughly equal to peak glottal flow in normal voicing. Fig. 11 shows the full-spectrum results.

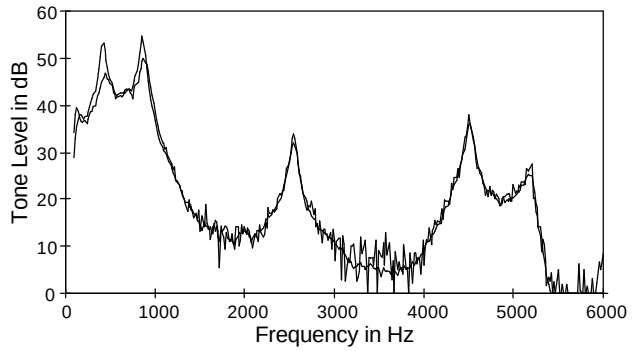

Figure 11. Frequency response of the mechanical vocal tract for no air flow (upper trace) and steady air flow of 860 ml/sec.

The effects of flow are more pronounced for the two lower formants than for the higher ones. Figs. 12 and 13 show expanded scales around the F1 and F2 peaks.

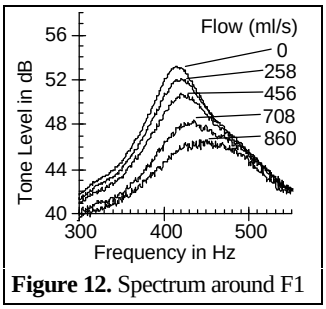

Figure 12. Spectrum around F1

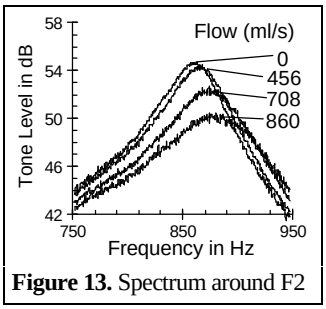

Figure 13. Spectrum around F2

Figure 14 shows the effect of flow on acoustic losses. Some of the bandwidth change could be an artifact of the instrumentation. Vorticity disperses sound wavefronts in direction and time, much as a rough surface scatters light. Time-phase jitter would lower the efficiency of a tone to pump energy into the resonance, and would reduce coherence of successive analysis sweeps. These effects would not alter the actual bandwidths as evidenced by rate of decay after a pulse.

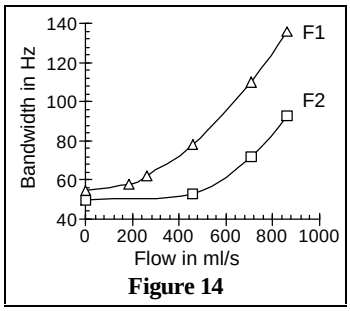

Figure 14

However, a given delay jitter maps into phase jitter proportional to frequency; vorticity cells are small compared to wavelength of the lower formants. Results of scattering should be smaller for low formants than for higher ones. Bandwidth changes from all causes were small for F3 and F4. Artifacts from scattering should definitely be smaller for F1 and F2 than the total changes for F3 and F4.

There are mechanisms for turbulence to produce true acoustic damping. Sound waves alternately speed up and slow the jet inside and coming out of a constriction. This in turn exchanges energy between sound and vortices in the shear layer. Since vorticity couples strongly to viscosity, much of the acoustic energy given up is not recovered.

Fig. 15 shows the effect of flow on formant frequencies. Resonances in a lossless uniform tube of length L would be odd multiples of F_1, where $1/F_1$ is the time for sound to traverse the length of the tube four times:

$$1/F_1 = 4L/C \;;\quad F_1 = C/4L \,.$$

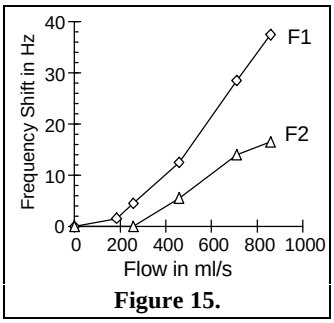

Figure 15.

With flow velocity V, the altered frequency $\tilde{F}_1$ becomes

$$1/\tilde{F}_1 = \frac{2L}{(C+V)} + \frac{2L}{(C-V)}; \qquad \tilde{F}_1 = \left(1 - V^2/C^2\right) F_1 \,.$$

For $V = .1\ C$ — Mach .1 over the full length — this would be a 1% downward shift. The actual result is an 8% upward shift for Mach .035 in the narrowest constriction, but much lower velocity over most of the tube. Clearly, there must be stronger effects, perhaps effects unique to non-uniform tubes.

The pattern of shifts — large positive shift of F1, smaller positive shift of F2 and essentially no shift of F3, F4 and F5 — is similar to the shifts produced by lowering the tongue body; that is, it is similar to an area change that enlarges the narrowest constriction, just forward of the velum. Of course, most flow effects go as particle velocity, and that is substantial only in small area. Thus we might look for effects that subtract from or otherwise alter the dominant acoustic term in small areas — the inertia term:

$$\frac{\partial p}{\partial x} = \frac{\rho}{a}\frac{\partial U}{\partial t}; \quad \text{acoustic inductance} \quad L_a = \frac{\rho}{a}.$$

For example, vorticity coupled to acoustics has a reactive component — effectively an inductance bridging L_a. One mechanism affects the area directly. Flow decreases thickness of the acoustic boundary layer; thus it enlarges the effective area, the portion of area not dominated by viscosity.

One goal of this study was to look for fluid-dynamic effects that might be significant in speech. The highest flow rate, 860 ml/s, is larger than the average glottal flow of ~250 ml/s, but not much higher than peak glottal flow, and peak flow inside the vocal tract. But, of course, such flows lead to particle velocities near Mach .03 only in the constricted voiced sounds /u, i, a/ and semivowels /w/ and /r/.

Bandwidth and frequency changes seen here are nearly an order of magnitude larger than perceptual JND's. Even so, steady-state shifts might not be discernible from small changes in area. But these shifts have a strong stochastic component; and the underlying flow is modulated by pitch and by the formants. JND's for frequency- and bandwidth-modulated formants are not known. In addition to the frequency/bandwidth shifts, flow produces noise — suppressed by this instrumentation, but probably audible.

The effects found here may well be perceptually significant. They are certainly the kind of monotony-breaking aperiodicity that could add to the impression of naturalness in real speech.