ABSTRACT

One part of our study is concerned with modelling of the respiratory system. A phonetically trained speaker uttered a sequence of similar sentences differing in stress patterns. Sub- and supraglottal pressures and oral flow were measured and from these data stylized articulatory trajectories representing glottal and oral constrictions were derived. A simplified mass, spring, resistance equivalent circuit of the pulmonary system was adopted for an analysis-by-synthesis derivation of underlying pulmonary forces. These amount to simple binary onset and offset commands for breath group initiation and offsets and a set of superimposed gaussian shaped pressure pulses for modelling stress patterns. Glottal and supraglottal articulations impose a passively induced fine structure on a $P_{sub}$ contour. This production oriented modelling supports the findings in speech analysis.

A second object of our study was the reading of a one minute long passage from a novel with the respiratory pressures included in the recording. A special listening test was carried out to grade the relative prominence of each syllable. We thus had the opportunity to add both production and perception data to the speech analysis. Of special interest was studies of the covariation of subglottal pressure with $F_0$ and their joint contribution to speech intensity measures. A general finding was that the subglottal pressure plays a role even at moderate stress level. It usually builds up to a maximum at the left boundary of a stressed syllable and decays with a rate that is positively correlated to the perceived prominence.

1. THE RESPIRATORY SYSTEM

1.1. Speech material

Our modelling of the respiratory system is based on data from a recording of sub- and supraglottal pressures together with the speech waveform, Fant, Hertegård and Kruckenberg [3]. A phonetically trained subject allowed his throat to be punctured and supraglottal pressure was sensed through a nasal probe.

Six short sentences of similar structure, differing in stress pattern were used: "Axell är här, Axel är här, År Axell här?, År Axel här?, År Axell här?, År Axel här?". The phonetic transcript of Axell is [ax'el] and of Axel [a'kel]. The sentences were recorded in sequence and our simulation includes pauses and inspiratory intervals between them.

Data on oral flow was needed for the modelling and had to be taken from a supplementary recording with a Rothenberg mask. Precautions were taken to make the subject produce a good match to the original utterance. Inevitable differences in tempo were corrected for with time warping by a dynamically varied resampling.

1.2. Estimating a pulmonary generator

Figure 1 shows the analogy used. Given the subglottal pressure $P_{sub}$ and the three pulmonary circuit elements we can then go on to compute what is the time development of the driving pressure generator

$$G = P_{sub} + M \cdot (dU/dt) + R \cdot U + K \cdot V$$

Knowing the oral flow $U$ we can estimate the flow time derivative in the sampled data sense as

$$(dU/d\tau)_0 = (U_0 - U_1)/\tau$$

and, from the flow integral, the lung volume

$$V_0 = V_1 - U_0 \cdot \tau$$

where 0 and 1 denote the current and preceding samples in intervals of $\tau$. Assuming G to be constant, then the R and M elements will cause decreases in $P_{sub}$ in proportion to the flow and its time derivative, respectively. With K being the elastic stiffness in the system, the KV term describes the discharge. At constant G, $P_{sub}$ will drop with time as the air escapes and the tension in the 'lung spring' is relieved in proportion to the lung volume decrease.

The MRK elements are constant for the whole speech material - they represent the anatomy of the speaker. Circuit element values have been suggested in the literature. Our modelling provided alternative data.
The resistive RU represents pressure drop in the pulmonar system and in the trachea. $P_{sub}$ is the output of the pulmonar generator with its internal RMK impedance. We disregard the impedances in the glottal and oral systems, varying with articulation. The combined effects of those are manifest in the known oral flow $U_{or}$. This is not precise - the analogy should properly be expanded several places with more K and M elements to account for e.g. compression in the pulmonary ($K_1$ in figure 1) and oral air volumes, and for inertia in the air passages. For the low frequency range considered here, below 100Hz, it was however found that this would be relatively insignificant.

Our intuitive approach was to select element values such that $G$ would come out with a representative shape, basically to select K such as to make $G$ grossly constant during those time intervals where the speech is going on. At inhalation the speaker will execute a large negative $G$ and tension the thoracical stiffness. During speech and other exhalations $P_{sub}$ is developed by the combination of this charged spring tension and whatever additional pressure the speaker cares to supply by muscular action.

The procedure was simulated with a spreadsheet program where $P_{sub}$, $U$, and $G$ were graphically displayed vs. time. Figure 2 shows the input data for a short excerpt, and the components which are added to $P_{sub}$ in order to develop the inferred $G$. It was found that the pulmonary mass $M$ did not contribute much, so it was set to zero. With $M$ we could get marginally better compensation of $P_{sub}$ transients at rapid flow changes, but this was obscured by the effects from residual pitch ripple in the $U_{or}$ waveform.

Then additionally, superimposed on this basically binary in- and exhalation scheme, one would expect some activity related to prosody, in particular there should be some indication of an intended stress pattern within a phrase. In this study $G$ was modelled as a sequence of Gaussian-shaped pulses, an arbitrary choice made from arithmetic convenience combined with visual similarity to the data tracks. Each such pulse is defined by three parameters for amplitude, width, and time location. Each sentence could be modelled with no more than six pulses (here denoted X0 through X5) where the first two were always assigned to the inspiratory phase. Then two or three pulses for the speech, and finally one pulse for expiration after each completed sentence.

![Figure 2: Traces for the fourth sentence in the sequence. The measured pressures and flow at bottom together with the assumed elements give the KV and RU traces above, inverted to highlight them as components of $P_{sub}$. Combining these with $P_{sub}$ makes the generator signal $G$ in the middle. $G$ is modelled as a sum $\Sigma X$ of 'lung pulses', also shown separately above. At top pitch and spectrogram for reference. X2 .. X5 do not coincide with the pitch track.](image-url)
2. THE PHONETIC FRAME

2.1. The breath group

Our acoustic-phonetic studies of prose reading are based on multiple parameter displays of the type illustrated in figure 3. In addition to oscillogram, spectrogram, F0 traces and intensity curves we have included a continuously scaled parameter of perceived syllable prominence, Rs, and when available also synchronous traces of sub- and supraglottal pressures, Psub and Psup. These allow a more complete interpretation.

The basic outline of the subglottal pressure contour within a declarative neutral sentence, see figure 3 left part, is characterised by an onset with a rise time of 100-150 ms up to an initial value of the order of 6-8 cm H2O followed by a decline of the order of 2-3 cm H2O towards the end which is approximately independent of the duration of the sentence.

The offset has a time constant of the order of 150 ms followed by a phase of rapid fall at a final abduction. The onset of voicing is associated with adduction and usually requires a subglottal pressure of 3-4 cm H2O but can be maintained down to 1 cm H2O. This is typical of a complete sentence or the initial clause of a sentence. A following secondary clause of a sentence starts with a somewhat lower Psub and shows less declination.

However, a Psub contour often departs from the neutral declarative norm. Local prominences of a focal character and more distributed rises or falls, in part supporting the intonation contour also appear, figure 3 right part.

The Psub contour also has a fine structure modulated by varying glottal- and supraglottal flow resistances. As a result, local minima appear at instances of glottal and supraglottal abduction, e.g. at transitions in and out of unvoiced sections or in intervals of breathy voicing such as in voiced [h], whereas Psub is restored during supraglottal closure.

The supraglottal pressure, Psup, varies within the extremes of zero at an open vowel and Psub at complete mouth and nasal closure. In our prose reading corpus we observe the following rank order of increasing supraglottal pressure: open vowels, closed vowels, nasals, [j], voiced [h], [l], [v], [r], [d], [g], [b], fricatives, stops. The rank order for subglottal pressure is similar but with rather small differences. An exception is the low Psub of voiced [h].

There was a noticable effect with respect to stress. Consonants were produced with on the average 1.5 cm H2O greater Psub in stressed than in unstressed syllables. In vowels the effect was limited to the order of 0.75 cm H2O.

2.2. Parameter interrelations

Problems of general interest we have been concerned with are the co-variation of F0 and Psub in speech and how they combine as determinants of intensity and sound pressure level, [2, 4, 6-8]. Although there are trends of a positive correlation, F0 is basically independent of Psub and is mainly determined by the cricothyroid muscle. It is known from earlier studies that a perturbation of Psub at constant laryngeal muscle activation causes a passive increase of F0. According to modelling performed by Titze [11], this effect is largely confined to low F0 phonations where the vocal folds are slack and lack stretching. Here the F0 increase is of the order of 4 Hz per cm H2O in Psub.

Titze [11] explains the F0 rise by an increase of the width of the glottal slit at constant length causing an elongation and

Figure 3: A declarative neutral sentence, left, and the beginning of a clause with a prominent focal accent, right.
stretching of the edge contour. However, this passively induced F0 rise is much smaller than observed from the average co-variation of F0 and Psub in connected speech.

A speakers available F0 range is about two octaves. A characteristic midfrequency for our male subject was F0r = 130 Hz. Fant et al. [8]. Most F0 data in connected speech are confined to frequencies up to F0r with intonation peaks extending into the upper range. We have found that Psub increases in proportion to F00.7 in the lower range, F0<F0r, and then tends to stay constant, except for singers who may show a continuing rise.

The sound pressure level, assuming a constant articulation, is basically set by the voice source excitation amplitude Ee, which is usually very close to the negative peak of the differentiated glottal flow, more precisely the flow derivative at the point of maximum discontinuity in the descending branch of the flow, Fant [2]. For frequencies up to F0r we find a joint contribution of F0 and Psub to Ee with a Psub 1.35 F0 1.35 proportionality. In terms of F0 alone, assuming normal coregation with Psub, this relation can be stated in the alternative form of Ee being proportional to F05. With F0 increasing above F0r, Ee reaches a maximum value, often as a peak followed by a decaying contour which stems from a contextually induced lowering of Psub. As a result, as F0 passes through F0r, there may appear a maximum of Ee, and at the peak F0 a minimum, [3], [7]. In other instances with a more stable Psub the Ee and thus the SPL reaches a saturation level in the F0 peak domain.

We have developed a model for predicting the SPL within the total range below and above F0r. With a 3 dB per octave F0 increase of SPL at constant Ee we arrive at
\[
\text{SPL} = K + 20 \log_{10} \left\{ P_{tr}^{1.1} x_1^{1.85} \left[ 1-x_2^2 + x_2^2/Q^2 \right]^{0.5} \right\} \text{ dB}
\]
where \( x_1 = P_{tr}/P_{tr}^{1.1} \) and Q=1.25.

Here the transglottal pressure, Ptr substitutes Psub. This equation was originally derived from glissando speech, Fant et al. [7], but has been successfully tested in connected speech.

### 2.3. Prominence correlates

Our continuously scaled prominence parameter Rs, Fant et al. [5], covers a range Rs=0-30. Typical values are Rs=10 for unstressed syllables and Rs=20 for stressed non-focally accented syllables. Linear regression analysis has provided the following approximate increases of acoustic parameters from Rs=15 up to the focal level of Rs=25.

Syllable duration, 125ms. Vowel duration of the order 80 ms. SPL 6 dB. SPL with high frequency pre-emphasis, SPLH, 9 dB.

A study of Psub showed a difference of the order of 1.5 cm H2O. This is close to what has already been noted in connection with the data on vowels and consonants. The Psub contour usually displays a build up towards the left boundary of a prominent syllable carrying a long vowel followed by a decay, thereby marking the location of the P-center in rhythmical analysis. The prominence Rs increases with the rate of Psub decline within the syllable. In focally accented syllables the F0 peak tends to occur within a domain of falling Psub as illustrated by the word “drog” in figure 3 right part. The relative latency of an F0 peak with respect to Psub has been observed by Collier [1].

A general conclusion is that subglottal pressure has a significant role not only in emphatic and contrasting stress but also at lower prominence levels. However, raised subglottal pressure does not appear to be a necessary requirement for a moderate focal accentuation in breath group final words.

A specific question is to what extent the F0-Psub dependency of F0 on Psub derived from Psub-F00.7, explains the observed mean declination of F0 within a breath group. The answer is that it occasionally is of the right order of magnitude but usually predicts a too large F0 fall. A better match would be the F0-Psub0.6 derived from Ladefoged [9; figure 12].

An additional phonetic function of subglottal pressure is in the realization of prosodic grouping, by modulating intensity onsets and offsets in boundary regions, even those without a proper pause. Our modeling of the respiratory system, developed in the first part of this paper, has derived a set of pulse shaped force functions underlying the prominence and grouping patterns we have observed in connected speech.

### 3. REFERENCES