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

A preliminary study of the consequences of intonation for the voice source

Pierrehumbert, J.

journal: STL-QPSR
volume: 30
number: 4
year: 1989
pages: 023-036

KTH Computer Science and Communication

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

A PRELIMINARY STUDY OF THE CONSEQUENCES OF INTONATION FOR THE VOICE SOURCE

*Janet B. Pierrehumbert**

Abstract

This paper reports the results of a pilot study on how intonational variables interact to determine characteristics of the voice source in English. By systematic variation of melody and voice levels within a sentence "Go from large up largest", it was possible to collect data on high and low F0 levels at 5 different voice intensity levels. Voice source parameters sampled from these conditions show systematic trends that do not in all respects conform with simple assumptions about underlying laryngeal behaviour and subglottal pressures. Thus, contrary to expectation, pulse skewing towards the right increased with F0 level at a constant voice effort. Pulse skewing also increased with overall voice effort. These phenomena are discussed with respect to general theory and previous investigations.

1. INTRODUCTION

Recent improvements in speech synthesis make it possible to envision practical applications in which extended passages of synthetic speech are used to convey complex information. However, many listeners still find the intonation of synthetic speech unacceptable; deficiencies are particularly evident in extended passages, where they can lead to poor comprehension of how different parts of the information are grouped and related to each other. This is the case even though algorithms have been developed for synthesizing fundamental frequency contours which closely resemble ones found in natural speech (Anderson, Pierrehumbert, & Liberman, 1984; Fujisaki & Hirose, 1984; Pierrehumbert & Beckman, 1988).

One contributing factor is the fact that current algorithms for synthesizing prosody map the underlying pattern into fundamental frequency alone, without attempting to model other differences in the voice source. The importance of such differences is illustrated in Fig. 1. This figure shows two narrow band sections for the vowel /a/ in the word "large", both spoken by the same speaker. The F0 value for both samples is 139 Hz. In one case, the word was spoken relatively high in the speaker's range at a low voice level (as if speaking quietly to a person close by). In the other, the speaker used a high voice level (as if projecting his voice across a large room), but because he used a different sentence intonation pattern, the word occurred towards the bottom of his range. In the first, the amplitude of the fundamental is 4 dB higher than the amplitude of F1. In the other, it is 10 dB lower. In the first, there is a 35 dB difference between the F0 amplitude and the amplitude of F3. In the second, this difference is only 10 dB. Fig. 2 shows estimates of the glottal flow derivative obtained by inverse filtering the same speech samples. The first sample has a much greater open quotient than the second and a more gradual return phase.

Failure to model such effects not only leads to unnatural voice quality. Probably, it also makes intonational differences less perceptible than in natural speech. Poor phonetic differentiation of different intonational categories may contribute to the impression that the informational organization is not well marked in synthetic speech.

---

*Dept. of Linguistics, AT&T Bell Laboratories (Current address: Dept of Linguistics, Northwestern University).

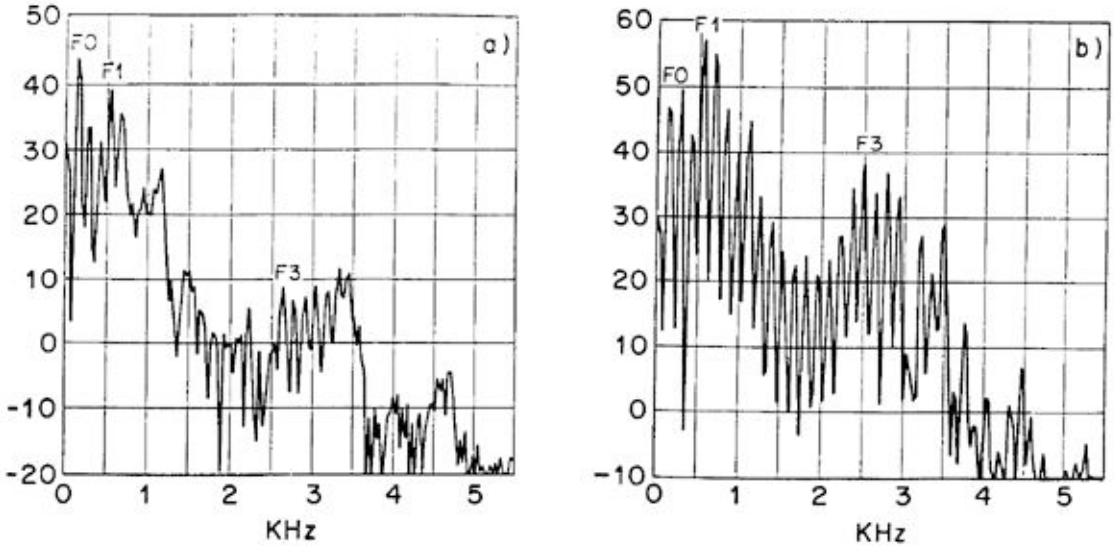

Fig. 1. *Narrow band sections for the vowel /a/ in the word "large" in two different utterances by the same speaker. The F0 value for both samples is 139 Hz. The utterances differed in their voice level and intonation pattern: a) relatively high in the pitch range, at a low voice level. b) relatively low in the pitch range, at a high voice level.*

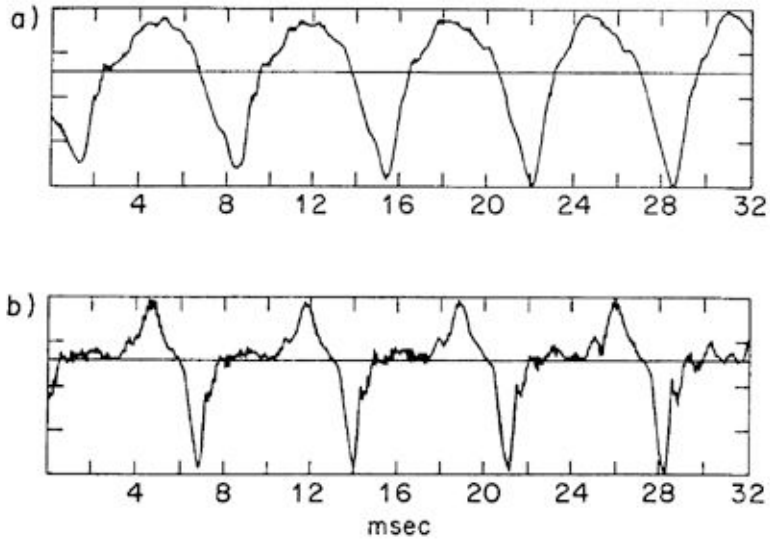

Fig. 2. *Estimates of the glottal flow derivative for the same two speech samples as in Fig. 1, obtained by inverse filtering: a) relatively high in the pitch range, at a low voice level. b) relatively low in the pitch range, at a high voice level.*

This paper reports the results of a pilot study on how intonational variables interact to determine characteristics of the voice source in English. The variables investigated were the melody and the voice level, or overall vocal effort. In order to motivate the experimental design, we give a brief summary of how these behave in English. In English, unlike Japanese or Swedish, any given sentence with a given stress pattern can be produced with many different melodies. The various melodies convey different messages about how the information in the sentence is related to that in other sentences and to the beliefs of the speaker and listener. Phonological analysis of the melodies reveals that there are particular places in the sentence where the speaker can exercise choice over the melody. At each of the most prominent

stressed syllables in the sentence, he can make a selection from the English inventory of "pitch accents." Beckman & Pierrehumbert (1986) list six different pitch accents, each a specification of a level in the overall pitch range or a movement between two levels. Three types are used in this study: low, high, and stepping. The data collected concerned the behaviour of the source at the pitch accent locations. A speaker also marks phrasal boundaries with additional melodic elements (Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986). The treatment of the phrasal boundaries varied in this study, but this variation was not analyzed, because only the pitch accent locations had appropriate segmental material.

The other dimension of variation in the study was voice level. Voice level, which is often viewed as peripheral to the linguistic system, nonetheless plays an important role in conveying the organization of extended texts in English as well as in other languages. The speaker raises his voice at the beginning of major groupings of sentences, and he adjusts his voice level phrase-by-phrase to highlight and background different information.

The consequences of phrasal voice level for F0 contours have been rather extensively investigated. In Pierrehumbert & Beckman (1988) and Liberman & Pierrehumbert (1984) phrases are taken to have an overall pitch range which controls how elements of the melody are realized as F0 target values. In the model developed by Fujisaki and his colleagues, see Fujisaki & Hirose (1984) and Fujisaki, Hirose, & Ohtas (1979), F0 values arise from the superposition of a phrasal contour with the contours contributed by local melodic elements. Although these approaches differ in many respects, they share the idea that phrase-level control of F0 contributes to the F0 value observed at any particular point in an utterance.

F0 is far from the only phonetic correlate of voice level. It has been known at least since the studies by Fant (1959) that raising or lowering the voice has very pronounced effects on the speech spectrum. These effects can be traced to the differences in the shape of the glottal waveform which arise as the subglottal pressure is varied, and also as different adjustments are made to the vocal folds. A raised voice is typically more pressed: the vocal folds are more adducted, giving rise to a shorter open quotient and a sharper closing than for a normal voice. These adjustments increase the amplitude of the high frequencies relative to the fundamental. In a lowered voice, the vocal folds are abducted, and they even may fail to achieve contact any time during the period. The amplitude of the fundamental is raised relative to the amplitude of the harmonics, and aspiration noise may become a significant component of the source.

Although it is well known that voice level affects these parameters in addition to F0, their role in the production of intonation has not been well studied. More data are needed in order to model these parameters in the same way that F0 has been modelled. In this study, an experimental design is applied which has proved to be valuable for the construction of F0 models: orthogonal variation of overall voice level and melody type. The linguistic phonetic structure is elucidated by examining how different melodies are produced at the same voice level, and how the same melody is produced in different voice levels.

In order to analyze experimental data and develop a model, a parameterization of the source is needed. Here, we use the Liljencrants-Fant (or LF) parameterization of the air flow through the glottis, which is described in Fant, Liljencrants, & Lin (1985). The model parameterizes the flow derivative rather than the flow itself. Pulse shapes for the flow derivative have two segments. The form of the first is given in Eq. (1), and the form of the second is given in Eq. (2):

$$U'(t) = E(t) = E_0 \cdot e^{\alpha t} \sin \cdot \omega_g t \qquad t_0 < t < t_e \qquad (1)$$

$$U'(t) = \frac{E_e}{\varepsilon \cdot t_a} \left[e^{-\varepsilon(t_c - t_e)} - e^{-\varepsilon(t - t_e)} \right] \qquad t_e < t < t_c \qquad (2)$$

In Eq. (1), E_0 is a scale factor, $\alpha = -B\pi$ where B is the "negative bandwidth" of the exponential, and ω_g is the underlying frequency for the pulse itself (ignoring the closed phase of the glottal cycle). In Eq. (2), E_e is an abbreviation for $E(t_e)$, and t_a is the time constant of the return phase following the maximum discontinuity. A condition of zero net flow gain within the period is usually imposed. The value of ε can be iteratively determined. Thus the model has four parameters in addition to the voice F0. Fig. 3, reproduced from Gobl (1988), shows an example pulse with various critical quantities indicated.

The two hallmarks of the model are the fact that the pulse shape has continuous derivatives from t_0 to the main excitation point t_e, and its provision for residual flow after this point. Since the DC component of the flow is not modelled, the portion of the flow covered by the second segment is called the "dynamic leakage." The earlier Liljencrants model also showed smooth behaviour between t_0 and t_e, but it had no provision for dynamic leakage. Dynamic leakage was incorporated into the model, Ananthapadmanabha (1984), after it was found to play a role in most real speech samples. However, the LF model uses an exponential curve to model the leakage instead of a parabolic one as Ananthapadmanabha (1984) proposed. This creates some additional complexity in fitting the model pulses to observed data, but yields a particularly simple spectral interpretation in terms of a first-order low pass filter.

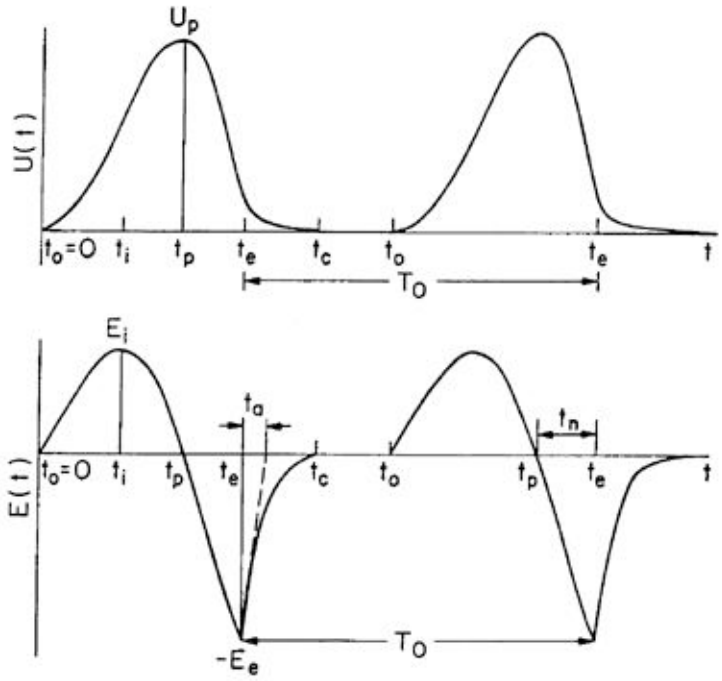

Fig. 3. *Glottal flow U(t) and the flow derivative E(t) with critical quantities indicated.*
U_p = peak flow.
E_i = maximum positive rate of change in the flow function.
E_e = negative level of the flow-derivative at maximum flow discontinuity (the excitation)
t_i, t_p, t_e = time points of E_i, U_p, and E_e respectively.
t_o = time of glottal opening.
t_c = time of complete (or maximum) closure.
$t_n = t_e - t_p$
t_a = index of dynamic leakage. (Figure is redrawn from Gobl (1988).)

The four parameters of the model can be computed from measurements of the inverse filtered speech. In practice, estimates of the beginning of the opening phase, the flow maximum (or zero- crossing in the derivative), the maximum negative excitation E_e, and the dynamic leakage are used to establish a fit. These points were all indicated in Fig. 3. It is important to understand that this and all other schemes for summarizing glottal flow can be mathematically reexpressed in many different forms. Thus, such models make no particular claims about which time or frequency domain measures are the most correct or insightful. Instead, their scientific contribution is at a more abstract level: they make claims about the number of significant dimensions of variation in the source, what constrains the relations among dimensions, and how time-domain measures are related to measures in the spectral domain.

A number of factors which may prove to be significant are neglected in this approach. There is no attempt to estimate the contribution of aspiration noise to the source. The contribution of subglottal poles and zeroes, which is probably significant for phonation modes with a relatively open glottis, is neglected. Acoustic interaction at the glottis, and all other nonlinearities, are ignored. The simplified approach is excused by the fact that it still captures many gross differences in the source, and thus promises to provide useful steps towards an improved understanding.

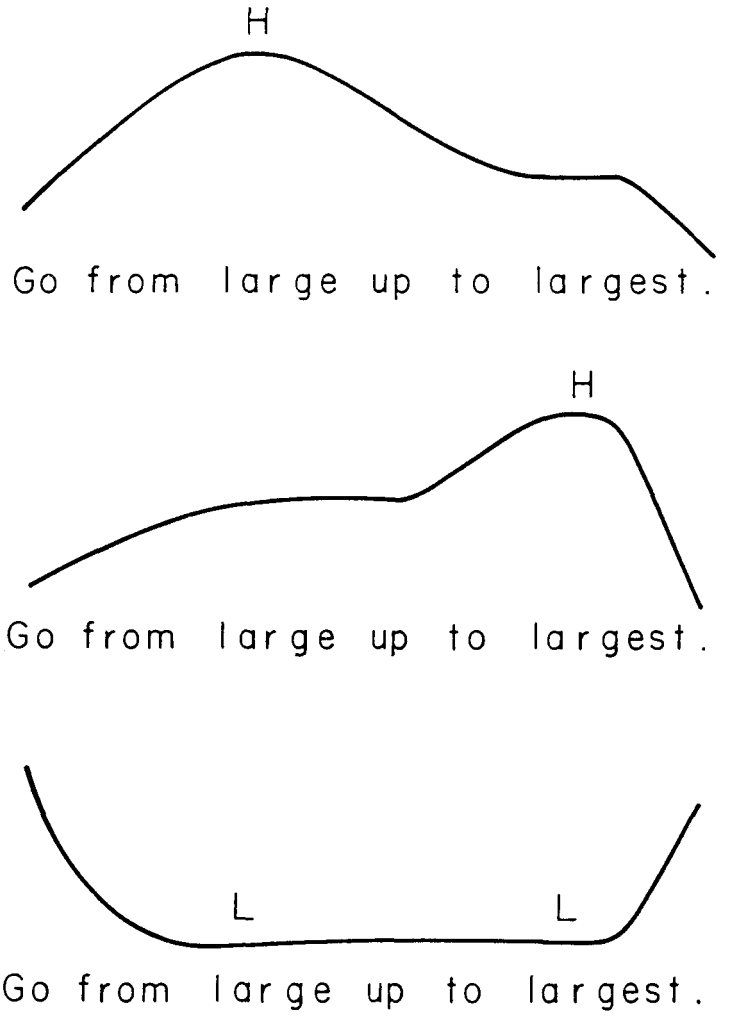

Fig. 4. Intonation patterns used in the experiment, as they were sketched for the subject: the "downstep pattern", the "declarative pattern", and the "contradiction pattern". "H" labels parts of the melody produced at the top of the pitch range for the utterance. "L" labels parts of the melody produced at the bottom of the pitch range for the utterance.

2. EXPERIMENTAL DESIGN AND PROCEDURES

2.1 Materials

Different renditions of the sentence "Go from large up to largest" were used in the experiment. This sentence was designed with three considerations in mind. First, the same syllable, "large" appears both early and late in the sentence, making it possible to look for positional effects. Second, the vowel in this target syllable, /a/, has a high first formant value, so that inverse filtering could be done more reliably than for some other sounds. Inverse filtering is a critical preliminary step to fitting the LF model in the time domain. Third, the semantics of the sentence allow it to be produced naturally with a variety of different intonation patterns.

Three intonation patterns were used in the study. All three had two pitch accents, one (the prenuclear accent) on the word "large" and the second (the nuclear accent) on the word "largest". In the "downstep pattern", the prenuclear accent appears as a fundamental frequency peak. The fundamental frequency steps down to an intermediate value at the nuclear accent, and then falls further to the bottom of the pitch range at the end of the sentence. In the "declarative" pattern, the fundamental frequency is higher in nuclear position than in prenuclear position. As in the downstep pattern, the F0 falls to the bottom of the range after the nuclear accent. In the "contradiction" pattern, both the prenuclear and nuclear accents are very low in the speaker's pitch range. The fundamental frequency rises after the nuclear accent rather than falling. The three patterns are sketched in Fig. 4, as they were on the cards which the subject read during the experiment. As is clear from the figure, the materials provided speech samples ranging from the bottom to the top of the speaker's current pitch range, in both the prenuclear and nuclear positions. In subsequent figures and discussion, "H" or "high" will be used to refer to data points from the prenuclear position in the downstep pattern and the nuclear position in the declarative pattern. "L" or "low" will be used to refer to either prenuclear or nuclear data points from the contradiction pattern.

The three intonation patterns were produced in five different voice levels. Voice level 3 was the normal voice level, and the others were defined in relation to it. Level 2 was "lower" and level 1 was "yet lower". Similarly, level 4 was "higher" and level 5 was "yet higher". In the subject's productions, level 1 appeared to be suitable for intimate conversation, while level 5 appeared suitable for projecting the voice in a noisy room.

2.2 Recording procedures

The speech was recorded in an anechoic chamber using a B&K microphone and a Sony PCM recorder. The recording set-up, which has been extensively used in acoustical studies of speech at the KTH and the University of Stockholm, is estimated to have an accurate frequency and phase response between 2 Hz and 4000 Hz. The subject attempted to keep his mouth a fixed distance from the microphone, so that sound pressure levels would be comparable across sentences. However, no head clamp was used.

The subject was a male native speaker of American English. Producing the fifteen variants of the sentence in randomized orders proved to be too difficult, so a fixed order was used. First the subject spoke in a normal voice, then in a lower voice, and then in an even lower voice. After this, he spoke above a normal voice and then in an even higher voice. For each voice level, the three intonation patterns were then produced in the order: downstep, declarative, contradiction. The speaker repeated the pattern if he felt he had produced it incorrectly. Ten repetitions of the utterance set were recorded.

Due to the great range of overall amplitudes produced, it was impossible to maintain a fixed setting of the amplifier and still have adequate dynamic range in the lowest voice levels. For this reason, the amplification was reduced by 10 dB in voice levels four and five. Subsequent processing of the measurements corrected for this adjustment.

2.3 Measurement procedures

The speech was digitized at 16 KHz with 12 bit accuracy, using a 6.3 KHz anti-aliasing filter. The gain was kept constant through the digitization session in order to assure comparability of measurements.

Digital wide-band spectrograms were made of all the tokens, which were used to establish time-points for analysis. The points selected were in the middle of the vowel in "large", just before the second formant began the rise induced by the following palatal. This point roughly coincides with the amplitude maximum of the vowel, and also precedes the drop in the third formant due to the following /r/.

Narrow band spectral sections at this point were used to determine the F0.

One set of fifteen variants was selected for the intensive analysis which is reported here. This was the fifth repetition of the set, which had been produced without any self-corrections by the speaker. At the analysis points in these tokens, the speech was first inverse-filtered using an interactive program written by Johan Liljencrants. Preliminary zero-phase high pass filtering at 20 Hz was used to remove slow variations due to air movements in the recording chamber. The inverse-filtering procedure itself, which is described in more detail in Gobl (1988), assumes 9 poles in the 0 to 8000 Hz frequency range of the speech. The researcher selects zeroes which, when applied as a filter to the speech, suppress peaks in the spectral domain and formant ringing in the time domain. Ordinarily, standard bandwidths are used for a preliminary fit, and the bandwidths are readjusted if noticeable discrepancies are observed. The zeroes are then interpreted as the resonances of the vocal tract, while the residue represents the glottal flow derivative. The signal can be integrated to give a picture of the glottal flow, but following Ananthapadmanabha (1984) and Fant & al. (1985) we view the flow derivative as being of more immediate descriptive interest.

The LF model was fit to the estimated flow derivative, using a program developed by Ananthapadmanabha. Again, the procedure is described in more detail in Gobl (1988). The researcher marks on the waveform the beginning of the opening gesture, the flow maximum (or zero-crossing in the derivative), the maximum negative derivative, and an index of the dynamic leakage. These points determine an LF fit, via the equations laid out in Fant & al. (1985). A model waveform and a spectrum are computed, and compared to the measured signal. The time points are readjusted if there are systematic errors either the time-domain or spectral matches.

The measures of the source which will be discussed here are EE, TA, OQ, and RK, as established by the model fit. EE and TA (E_e and t_a) are as shown in Fig. 3. EE is a measure of the excitation strength. The absolute value is reported. TA is a measure of the dynamic leakage. OQ, the open quotient, is defined as $(t_e-t_0)/T_0$; that is, only the pulse up to the major discontinuity is considered to be "open", and the dynamic leakage is neglected. This measure has been found to behave more similarly to the traditional understanding of open quotient than a measure including the leakage (Gobl, 1988). RK, defined as $t_n/(t_p-t_0)$, is a measure of glottal pulse symmetry. It takes a value of 1.0 when the opening and closing segments of the U(t) pulse have the same duration. Values above 1.0 increase in relation to leftward skew. Values below 1.0 decrease in relation to rightward skew. In defining RK, the dynamic leakage is excluded from the closing segment of U(t) so that the measure will have a well-behaved relationship to OQ and to traditional ideas of pulse symmetry.

All of these measures could in principle be estimated by single point measurements of the inverse-filtered signal. Going through the stage of model fitting allows all points in the signal to contribute to the values. This stabilizes the measures statistically, particularly in cases where there is significant noise or interaction ripple. In these cases, the beginning of the opening phase and the dynamic leakage can be quite difficult to estimate visually. The model fit also supports a translation between the time and spectral domains. This is important, be-

cause in some cases the differences among alternative interpretations are much more salient in the spectral domain than in the time domain.

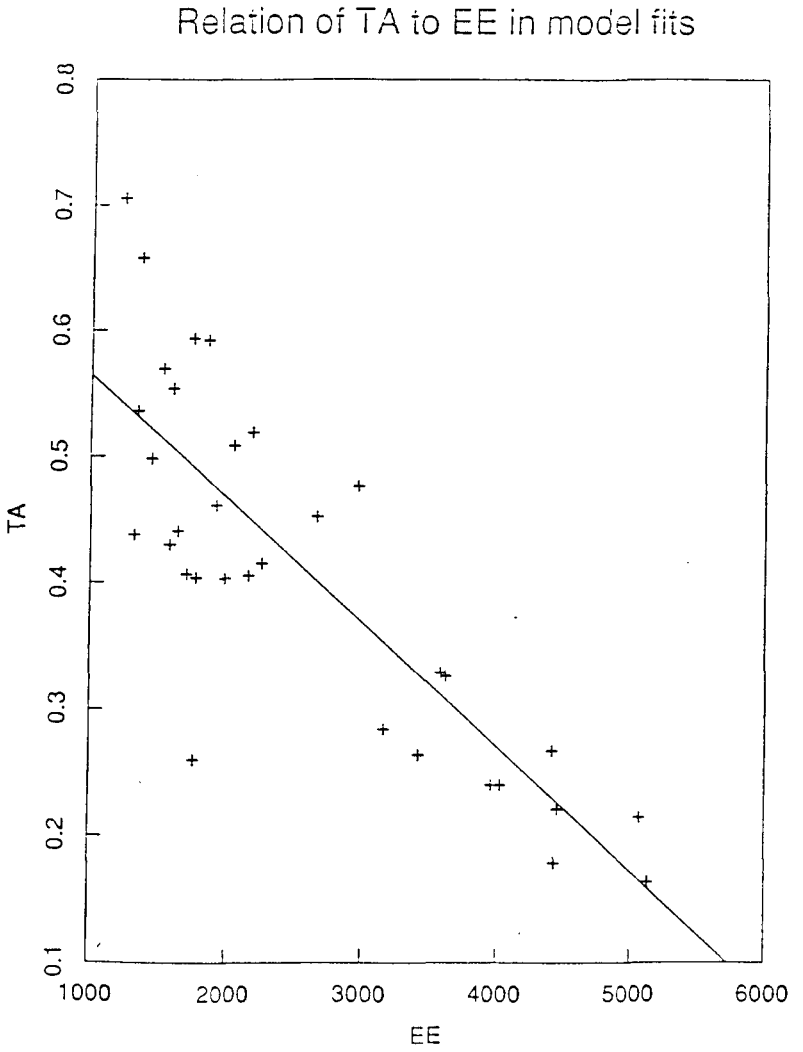

Fig. 5. Relation of TA to EE in model fits, with a summary regression line.

3. RESULTS

As expected, the F0 values for H's were higher than for L's, in all voice levels. The size of the difference between H and L increased with voice level, chiefly because the H's were higher. The H tone values ranged from 140 Hz at voice level 1 to 225 Hz at voice level 5. The L values also went up with voice level, from 91 Hz in level 1 to 125 Hz in level 5. This outcome is in superficial contradiction to the result of Liberman & Pierrehumbert (1984), that the low value at the bottom of the range (the "baseline") is invariant for each speaker across changes in voice level. However, the measurements in Liberman & Pierrehumbert (1988) were made at the very end of the utterance whereas the measurements in this experiment were all made utterance-medially. The outcomes can be reconciled by assuming that whatever adjustments the speaker makes to raise his voice level are relaxed by the time he finishes speaking.

F0 alone was not a good predictor of source characteristics. This was the case because F0 is only a one-dimensional parameter, and the source data showed structure relating to two dimensions: overall voice level, and relative location in the range established by the voice level. We bring out this point by considering the behaviour of each of the source parameters EE, TA, OQ, and RK as a function of voice level and tone type.

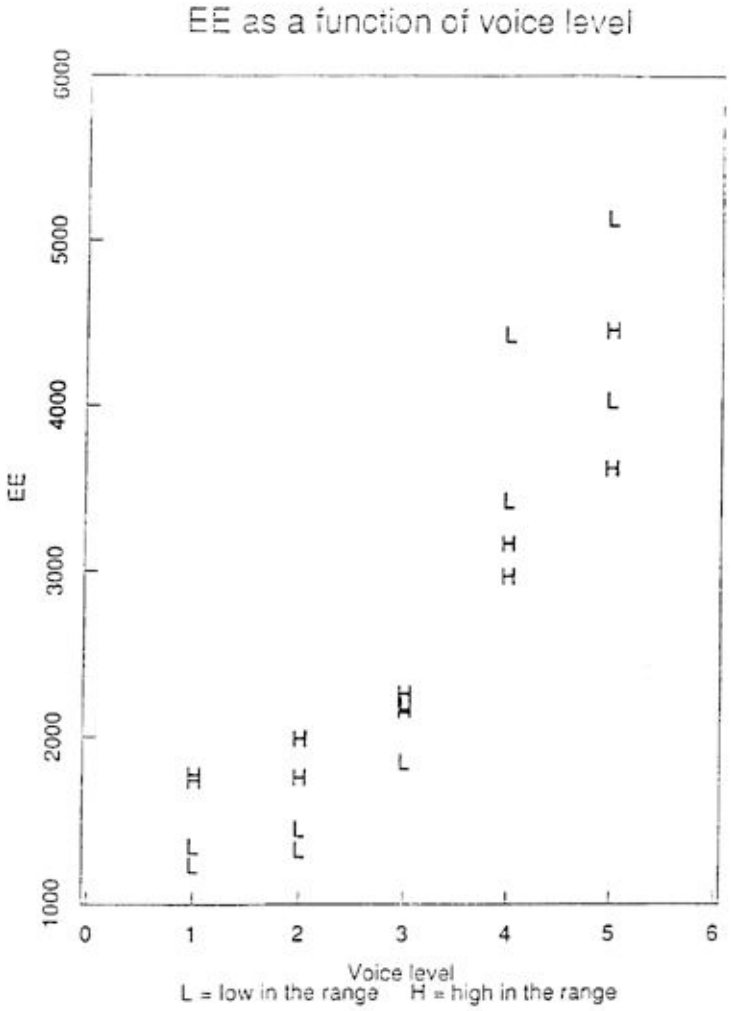

Fig. 6. EE as a function of voice level, separating H and L values.

Spectral computations based on the LF model show that EE and TA are the main determinants of the formant levels. EE controls the overall spectral level above the region of the fundamental, while TA differentially affects the formants by imposing a first-order low-pass filter. These dimensions of variation are certainly well-separated from a mathematical or perceptual point of view, but they appear not to function independently in production. Fig. 5 shows the relation of TA to EE in the model fits, with a summary regression line. The values are highly negatively correlated. A more detailed breakdown of the data, not shown here, indicated that the difference between H and L made no further contribution to explaining the statistical variability in TA values. A heuristic argument relating EE and TA is made in Gobl (1988), which investigated effects of stress and segmental type on the voice source. It is interesting to see that the same result obtains under the rather different conditions of variation investigated here.

The value of EE is largely determined by the voice level, as shown in Fig. 6. However, there is a small interaction of tone type with voice level, as indicated by the plotting characters. At low voice levels, H tones had slightly higher EE values than L tones, while at high voice levels the relationship was if anything reversed. As a result, the total range of EE values for H tones was less than for L tones.

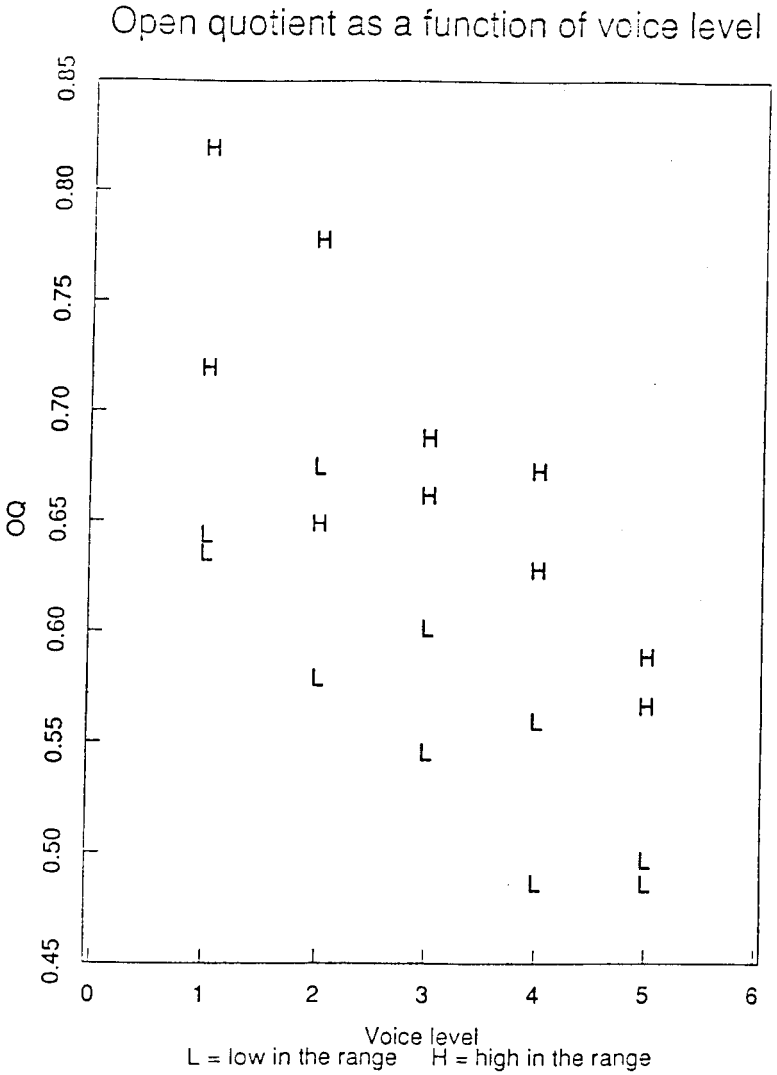

Fig. 7. OQ as a function of voice level, separating H and L values.

In Fig. 7, the open quotient OQ is plotted as a function of voice level and location in the range. Overall, H tones have a greater open quotient than L tones, and higher voice levels have a lesser open quotient than lower voice levels. Since cricothyroid contraction is the main mechanism for the production of H tones, they involve greater vocal fold tension than L tones. From this fact, both the Ishizaka-Flanagan and the more elaborated Titze-Talkin models of vocal fold vibration predict that they will show a greater open quotient. The decrease of open quotient with voice level can be explained in these models by assuming an increase in subglottal pressure. (See computed flow patterns in study by Monsen, Engebretson, & Vemula (1978) of the Ishizaka & Flanagan (1972) model, and in the Titze & Talkin (1979) model. In studying this figure it is important to bear in mind that either increasing the voice level or going upwards in the pitch range increases the F0. Thus, F0 can be seen not to have uniform consequences for OQ; the OQ can only be predicted by looking at the underlying linguistic reasons for the particular F0 value.

The symmetry measure RK is plotted as a function of voice level and location in the pitch range in Fig. 8. All tokens have RK values less than 1.0, indicating rightward skew of the pulse shape, as is typically observed. In all voice levels, H tones have lower RK values than low tones, indicating greater skew to the right. This result is surprising, since in both the

Ishizaka-Flanagan (1972) and Titze-Talkin (1979) models the higher vocal fold tension responsible for F0 raising results in a more symmetric pulse shape. Computer flow patterns illustrating this prediction can be found in Monsen & al. (1978) and Titze & Talkin (1979).

Examination of the LF fits to the inverse filtered speech indicated that this result was not an artifact of the fitting procedure. On the contrary, the extent of rightward skew for L tones was systematically overestimated, especially in the lowest three pitch ranges. This happened because the pulses in U'(t) typically had a shoulder in the region of the negative zero-crossing which was not well described by the LF model; the LF model assigned the negative zero-crossing to the right of its true location in order to accurately approximate the immediate region of the closure. The corresponding observation in the U(t) domain is that the LF model had difficulty in fitting the observed combination of a broad peak and a sharp closing gesture, and compromised by assigning the maximum towards the right of the true peak location. LF fits to the inverse-filtered H tone examples did not appear to exhibit any systematic errors.

The decrease in skew in Fig. 8 thus indicates that increasing the subglottal pressure is not the only mechanism involved in raising the voice. To explain the data, it appears that some laryngeal adjustments must also be involved. Any estimate of these adjustments is constrained, however, by the data on the behaviour of the open quotient; different strategies would carry different predictions about the behaviour of the open quotient. Detailed computations with a physiological model are thus necessary to reconcile these data.

The discrepancy between the present data and the laryngeal models leads us to speculate that the laryngeal models may not well describe the laryngeal state which occurs when the F0 is actively lowered for a stressed L tone. In the mid to high region of the pitch range, the dominant mechanism for F0 control appears to be cricothyroid contraction and relaxation. This mechanism, as it translates into vocal fold tension, may well affect pulse shape as described in the models. For example, our own informal observations suggest that the extremely high F0 values found at the end of yes/no questions exhibit more symmetric pulse shapes than the H tones in this study, whose associated F0 values were far lower. However, the mechanism for producing stressed L tones, which are presumed to be produced below the mid region of the pitch range, appears to be poorly understood. A combination of muscular actions may be involved. More physiological understanding would be needed in order to assess the applicability of present laryngeal models.

We also find that the RK decreases, i.e., that skew decreases as voice level increases. The physiological interpretation of this observation is also complex. Increasing the voice level is probably accomplished by adjustments to the vocal folds as well as an increase in subglottal pressure. As Fant (1982) demonstrates, an increase in subglottal pressure will decrease the skew, provided that the glottal area function is not changed. As the subglottal pressure increases, the combined glottal vocal tract and inductance becomes a less dominant determinant of the pulse shape, and so the flow pattern tracks the area function more closely. However, models of vocal fold vibration show that increasing the subglottal pressure has a substantial effect on the shape of the glottal area function. At increased pressures, the area function is much more skewed, apparently more than offsetting the reduced role of the inductance. As a result, computed flow patterns show skew increasing with subglottal pressure (Monsen & al., 1978; D. Talkin, personal communication).

The qualitative behaviour of RK and OQ are thus consistent with the conclusion that increased subglottal pressure is a major mechanism in raising the voice level. However, increasing the vocal fold adduction also results in a smaller open quotient and a more skewed pulse shape. Therefore, detailed computations to establish the quantitative correspondence between empirical data and the physiological theory would be necessary in order to establish whether adjustments to the vocal folds are also critically involved in control of voice level.

4. DISCUSSION AND CONCLUSIONS

The data presented here show that location in the range and voice level affect source parameters differently. Given these results, F0 alone cannot be expected to provide an adequate phonetic realization of intonation in synthetic speech. The qualitative behaviour of all parameters discussed is consistent with the idea that subglottal pressure is a major mechanism for control of voice level, but vocal fold adduction may also be critically involved. A surprising contrast was found between OQ, which increases the F0 at each voice level as predicted by physiological modes, and RK, which decreases with F0 contrary to prediction. It is suggested that an improved understanding of active F0 lowering is needed to explain this discrepancy.

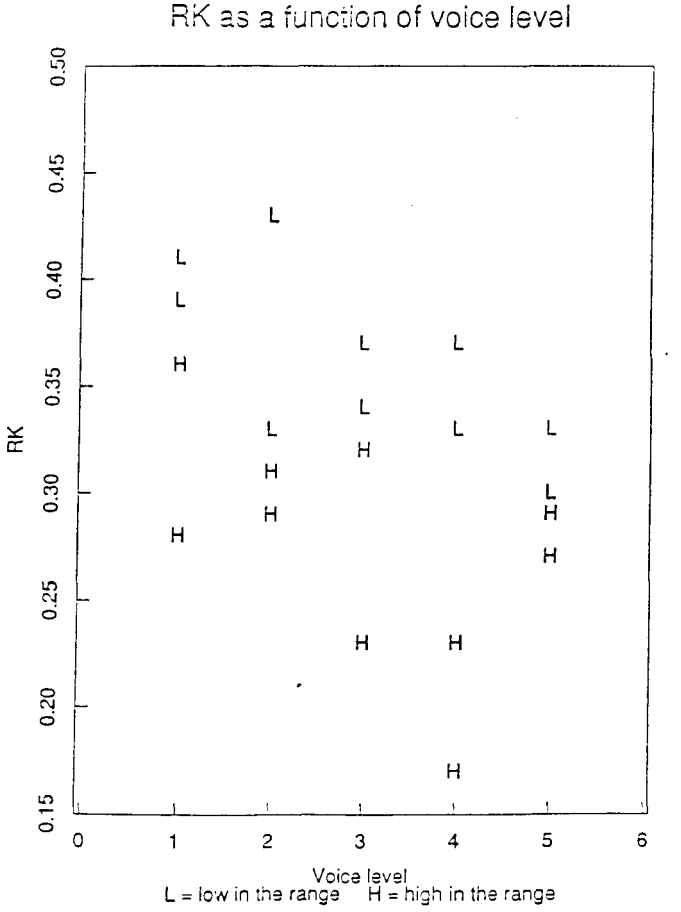

Fig. 8. RK as a function of voice level, separating H and L values.

The very interesting previous work on English intonation by Monsen and colleagues and (Monsen & Engebretson, 1977; Monsen & al., 1978) also established that describing the consequences of different linguistic intents requires reference to a description of the source which is at least two dimensional. By examining transitions in overall spectral shape and the relation of F0 to RMS, they identify separate contributions of vocal fold tension and subglottal pressure to the phonetic expression of intonation. Although their experimental design contrasted declarative and interrogative intonation patterns, presumably exhibiting a contrast between stressed H and L tones, we are not able to deduce from their report whether or not their results on the relative skew for these tones were similar to ours. Their measures do not clearly distinguish effects on open quotient from effects on pulse symmetry, and while they report contrasts in the type of spectral transitions observed within each utterance, they do not provide comparisons of spectral characteristics for comparable locations across utterance types. The generality of Monsen & al.'s results is limited by the fact that they obtained the source function with a

Sondhi tube (1975), which constrained them to use materials comprised entirely of neutral vowels. In addition, an informal handling of the English intonation system resulted in an experimental design which confounded tone and stress. In spite of these limitations, the study remains a valuable effort to relate source characteristics to the speaker's communicative intentions.

More recent and sophisticated work on the description of the source increases the number of dimensions of variation which might be conjectured to function in communication. The materials in the present experiment are not well suited to placing an upper bound on the number of independent dimensions of variation. The experimental design was three dimensional, and one dimension (the prenuclear-nuclear distinction) was not a strong candidate for being related to an independent physiological dimension. In particular, the materials do not allow us to separate the roles of ligament stress and vocalis stress in the Titze-Talkin model. Nonetheless, the study makes a small contribution in this direction by supporting Gobl's suggestion that parameter TA of the LF model is predictable from EE. Further experiments in which the linguistic factors controlling the source are orthogonally varied will be needed to elucidate this issue fully.

5. ACKNOWLEDGEMENTS

This work was carried out at the Royal Institute of Technology, Stockholm, under support by AT&T Bell Laboratories and the NSF US-Sweden Cooperative Science Program. I am very grateful to the Department of Speech Communication and Music Acoustics for the use of their facilities, and to Prof. Gunnar Fant, Christer Gobl, and Qiguang Lin for their advice and assistance. I would also like to thank David Talkin, Ann Syrdal, and Kim Silverman for their comments on the manuscript.

References

Ananthapadmanabha, T.V. (1984): "Acoustic analysis voice source dynamics," STL-QPSR 2-3/1984, pp. 1-24.

Anderson, M.D., Pierrehumbert, J., & Liberman, M.Y. (1984): "Synthesis by rule of English intonation patterns," *Proc. IEEE ICASSP*, pp. 2.8.2-2.8.4.

Beckman, M. & Pierrehumbert, J. (1986): "Intonational structure in Japanese and English, pp. 255-310 in *Phonology Yearbook 3*.

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

Fant, G. (1982): "Preliminaries to analysis of the human voice source," STL-QPSR 4/1982, pp. 1-27.

Fant, G., Liljencrants, J., & Lin, Q. (1985): "A four-parameter model of glottal flow," STL-QPSR 4/1985, pp. 1-13.

Fujisaki, H. & Hirose, H. (1984): "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J.Acoust.Soc.Japan **5**, pp. 233-242.

Fujisaki, H., Hirose, H., & Ohta, K. (1979): "Acoustic features of the fundamental frequency contours of declarative sentences in Japanese," Ann.Bull. **13**, R.I.L.P., University of Tokyo, pp. 163-173.

Gobl, C. (1988): "Voice source dynamics in connected speech," STL-QPSR 1/1988, pp. 123-159.

Ishizaka, K. & Flanagan, J.L. (1972): "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Syst.Tech.J. **51**, pp. 1233-1268.

Liberman, M.Y. & Pierrehumbert, J. (1984): "Intonational invariance under changes in pitch range and length," pp. 157-233 in *Language Sound Structure: Studies in Phonology Presented to Morris Halle* (Aronoff, M. & Oehrle, R.T., eds.), MIT Press, Cambridge.

Monsen, R. B. & Engebretson, A.M. (1977): "A study of variations in the male and female glottal wave," J.Acoust.Soc.Am. **62**, pp. 981-993.

Monsen, R.B., Engebretson, A.M., & Vemula, N.R. (1978): "Indirect assessment of the contribution of subglottal air pressure and vocal-fold tension to changes of fundamental frequency in English," J.Acoust.Soc.Am. **64**, pp. 65-80.

Pierrehumbert, J. (1980): "The phonology and phonetics of English intonation," MIT Ph.D. diss.; available from Indiana University Linguistics Club, Bloomington, Indiana.

Pierrehumbert, J. & Beckman, M. (1988): *Japanese Tone Structure, Linguistic Inquiry Monograph 15*, MIT Press, Cambridge.

Sondhi, M. (1975): "Measurement of the glottal waveform, J.Acoust.Soc.Am." **57**, pp. 228-232.

Titze, I.R. & Talkin, D.T. (1979): "A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation," J.Acoust.Soc.Am. **66**, pp. 60-74.