The use of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal, and loud phonation

Paavo Alku¹, Juha Vintturi², Erkki Vilkman³
1: Helsinki University of Technology, Finland
2: Helsinki University Central Hospital, Finland
3: University of Oulu, Finland
Paavo.Alku@hut.fi

Abstract
A method is presented to estimate the effect of intentional raising of fundamental frequency (F0) on vocal intensity. The method, Energy of the Synthesised Period (ESP), is based on computation of the energy of a hypothetical speech sound synthesised using a single period of the glottal volume velocity waveform and a digital filter that models the vocal tract. Both the glottal flow and the vocal tract filter are estimated by inverse filtering. The results show that, in producing loud voice, speakers use F0 to increase the number of glottal closures per time unit, which increases rapid fluctuations in the speech pressure waveform, which, in turn, raises vocal intensity. The average increase of sound pressure level due to this active use of F0 was approximately 4 dB in loud speech.

1. Introduction
In speech science, the term "(vocal) intensity" is commonly used to refer to the acoustic energy of speech, and it is typically quantified by using the sound pressure level (SPL) [1]. According to Titze [1], there are three basically distinct mechanisms to control vocal intensity. They correspond to adjustments of the vocal apparatus below the larynx, within the larynx, and above the larynx. Below the larynx, intensity can be regulated by controlling the aerodynamic output of the lungs to the vocal system. Within the larynx, there are methods to regulate intensity by modifying the vibration of the vocal folds and hence changing the amount of aerodynamic power converted into acoustic power. These methods correspond, for example, to increasing the maximum flow amplitude or to decreasing the length of the glottal closing phase. Above the larynx, vocal intensity can be modified by changing the formant settings of the vocal tract. This implies, for example, adjusting the resonances of the vocal cavity to coincide with the harmonics of the glottal source.

A number of investigations were made in the late 50's and early 60's on the intensity regulation of speech [2-4]. These studies addressed mainly the impacts of subglottal pressure and air flow rate on vocal intensity. Later, when the use of inverse filtering became popular, numerous experiments were conducted to study the relationship between the characteristics of the glottal volume velocity waveform and vocal intensity [5-8]. From these studies it is known that, among the different voice source parameters, the one that is most closely related to vocal intensity is the negative peak amplitude of differentiated glottal flow [6, 9].

Subglottal pressure or, rather, lung pressure, is known to be one of the key factors in the regulation of speech intensity [10, 11]. Raising of subglottal pressure increases vocal intensity because it amplifies the peak flow of the glottal pulse and thereby also the negative peak amplitude of differentiated glottal flow [8]. An increase of subglottal pressure does not only affect the amplitude features of an individual glottal pulse but also contributes to the rate of repetition of consecutive glottal pulses, i.e., the fundamental frequency (F0) of voice. Many studies have been published on the relationship between F0 and vocal intensity (or loudness) [7, 12, 13]. In one of their studies, Gramming et al. [12] reported that the mean pitch increased by about a half-semitone when intensity was increased by one decibel. According to [12], the increased value of F0 can be considered a passive result of the raising of subglottal pressure in order to produce louder sounds. However, the intentional use of F0 may also have an important role in the regulation of speech intensity when formant tuning is used [1]. This implies that F0 of the voice is adjusted so that one of its lowest harmonics coincides with a resonance of the vocal tract, typically the first formant (F1).

As pointed out by Gramming et al. [12], the raising of subglottal pressure in order to create louder voice causes a passive increase of fundamental frequency. On the other hand, it is also possible that F0 is used actively to produce loud voice because an increase of F0 results in a larger number of speech pressure cycles per time unit, and SPL is hence raised. Therefore, our first rationale was to present a method which would allow a quantitative analysis of the role of F0 (i.e., passive vs. active) in the regulation of speech intensity. The second rationale for the present study was the need to shed light on the relationship between F0 and SPL in circumstances where speakers are allowed to use their fundamental frequency freely to produce voices of greatly different intensity levels.

2. Material and methods

2.1. Speech material
Speech data were collected from eleven adult Finnish speakers (five females) with no history of speech, voice, or hearing disorders. Each subject produced a series of the word /pa:ppa/ by gradually increasing loudness. The acoustic speech pressure waveform was recorded using a condenser microphone (Brüel&Kjaer 4176) placed at a distance of 40 cm from the lips of the speaker. For the computation of SPL-values, we also recorded a calibration signal generated by a Brüel&Kjaer 4231 calibrator.

The subjects repeated the /pa:ppa/ words by increasing the SPL-values in approximately 5 dB steps from the softest voice up to the loudest, with an SPL-value of 105 dB. Some subjects voluntarily also produced the loudest sound with an SPL-value of 110 dB.) The first phonation sample was to be...
produced as softly as possible without whispering. The output level of the speech signals was controlled by means of a sound level meter (Brüel & Kjær 2225). In order to be able to increase the output level of speech in approximately 5 dB steps, the subjects watched the LED light display of the sound level meter and they gave multiple productions until the desired SPL increase from the previous token was achieved. The subjects were given no other restrictions regarding their voice production, which means that pitch and phonation type were chosen freely by the speakers during the recording. In order to collect enough data representing high vocal intensity, the subjects repeated the three loudest speech samples (SPL-values 95 dB, 100 dB, and 105 dB) three times. The total number of speech samples produced was 86 and 102 by female and male speakers, respectively.

Signals were transmitted to a computer via a high quality audio card using a sampling frequency of 22050 Hz. In the computer, the speech waveforms were high-pass filtered in order to remove any possible low-frequency air pressure variation picked up during the recordings. Finally, SPL-values were determined for all the speech signals over the long /a/ vowel using energy computation and the SPL-value of the calibration tone (94 dB) as follows:

$$\text{SPL}_{\text{speech}} = 94\text{dB} + 10\log_{10}\left(\frac{\text{Energy}_{\text{speech}}}{\text{Energy}_{\text{calibration}}}\right) \tag{1}$$

2.2. Inverse filtering

In order to estimate the glottal volume velocity waveforms, we used an inverse filtering technique that is a slightly modified version of the technique described in [14]. The modification concerned modelling of the vocal tract transfer function. In the present study, the estimation of the vocal tract transfer function was based on Discrete All-Pole Modeling (DAP) [15] instead of the conventional Linear Predictive Coding (LPC), which was used in [14]. The estimation of glottal flow was computed by using an analysis window of 50 ms. (For some of the low-pitched male voices, the length of the time window was increased to 70 ms in order to cover at least four glottal cycles.) In order to estimate the glottal flow over sustained phonation, the analysis window of inverse filtering was set at the middle of the long vowel in the /pa:ppa/ word.

2.3. Energy of the Synthesised Period (ESP) - measurement of vocal intensity by excluding the effect of fundamental frequency

Computation of SPL of a speech sound, as shown by Eq. 1, is based on measuring the energy of the signal over a certain time-window. This computation takes into account the speech signal over several glottal periods. Consequently, SPL will be affected by the features of a single glottal pulse (e.g., AC-amplitude and slope of the closing phase) as well as by the characteristics of the vocal tract (e.g., formant bandwidths) but, importantly, also by F0. In order to find out how speakers use F0 as a method of intensity regulation, we developed a new parameter, Energy of the Synthesised Period (ESP). The idea in deriving this parameter was to measure how much energy can be generated by the resources of a single glottal pulse together with vocal tract filtering. However, the contribution of several glottal periods (i.e., the effect of F0) on the sound energy was to be excluded. If vocal intensity is regulated by changing either the features of the single glottal pulse or the characteristics of the vocal tract, the values of SPL and ESP would show an equal change. However, if F0 is used in order to raise vocal intensity, only SPL would become larger. Hence, to follow ESP as a function of SPL when changing phonation from soft to loud would make it possible to analyse how F0 raising is used as a strategy for increasing intensity.

The computation of ESP is based on the source-filter theory of speech production [16]. According to this theory, a voiced speech signal is produced as a combination of the following three cascaded processes: the glottal excitation, the vocal tract filtering and the lip radiation effect. In order to exclude the effect of F0, the ESP method first separates the given speech signal into the three processes mentioned above. Then, a hypothetical speech signal is synthesised by re-combining the three separated processes so that only a single glottal cycle is used from the excitation process but the rest of the features of the processes are left intact. This synthesis is done by estimating, firstly with inverse filtering, the glottal flow waveform and the all-pole model of the vocal tract. Secondly, one glottal cycle is cut from the middle of the glottal flow waveform obtained. The hypothetical voice signal is then synthesised by using this single glottal pulse as an excitation to the all-pole vocal tract model given by inverse filtering and by differentiating the output in order to take into account the lip radiation effect. Finally, the value of ESP is obtained by computing the energy of the hypothetical signal, denoted in the discrete time domain by s(n), as follows:

$$\text{ESP} = 10\log_{10}\sum_{n=0}^{\infty} s^2(n) \tag{2}$$

In principle, the duration of the hypothetical signal is infinite due to the application of all-pole filtering in modelling of the vocal tract. However, since the all-pole filter given by inverse filtering is stable, its response will attenuate rapidly towards zero. Therefore, the energy of the hypothetical signal given in Eq. 2 can be computed over a finite time span. In the current study, we used the time span of 50 ms in this energy computation.

It is worth noticing that the energy computation in ESP is performed over a hypothetical signal generated using the source-filter theory, which is not the same as determining the energy over a single cycle of the original speech signal. This comes from the fact that the latter is based on processing the output of the vocal apparatus, which has been excited by a train of consecutive glottal pulses. Since responses of individual pulses will superimpose, the energy computed over a single cycle of the original speech waveform can be affected by the previous fundamental periods (i.e., the energy is affected by F0). Once this superimposition has occurred it is not anymore possible to measure from the original speech waveform, to what extent the energy computed over one cycle has been affected by fluctuations from the previous periods.

3. Results

The obtained values of SPL, computed using Eq. 1, varied between 54 dB and 106 dB in the voice samples produced by the female subjects and between 57 dB and 111 dB in the samples produced by the males. Since there was a large dynamic variation in the SPL-values, the speech samples were divided into three intensity categories: soft (SPL≤70 dB), normal (70 dB≤SPL≤90 dB), and loud (SPL>90 dB).
Figure 1 shows an SPL-ESP-graph computed from voice samples of a female speaker. As can be seen from the figure, the value of ESP parallels SPL in a manner close to linear. It is worth noticing that Fig. 1 has been drawn by using the same range of 70 dB on the x-axis as on the y-axis. This implies that if ESP follows SPL as a linear function whose slope equals unity (i.e., $\text{ESP} = 1.0 \cdot \text{SPL} + \text{IN}$, where IN denotes the intercept of the linear function) then the ESP-SPL-graph should form a line at an angle of 45 degrees to the x-axis. This line, denoted as $\text{ESPlin}$, is depicted in Fig. 1. In accordance with the definition of ESP presented in section 2.3, this linear relationship implies, in turn, that the raising of the energy of the synthesised single period is paralleled by an exactly equal increase of SPL. However, a raise in vocal intensity that causes a larger increase in the SPL-value than in the ESP-value implies that there must have been an additional source of energy present in the computation of SPL, but absent from the computation of ESP. According to the definition of ESP, the only such factor is fundamental frequency. Hence, the deviation of the ESP-SPL-graphs from $\text{ESPlin}$ implies that the speakers have used intentional raising of F0 as a method to increase vocal intensity. (The intercept of $\text{ESPlin}$ was optimised with the mean squares error criterion among the speech samples with SPL-values below 70 dB. It should be noticed that this optimisation did not correspond to the computation of linear regression because the slope of the line was fixed to unity prior to optimisation. The obtained values for the optimal intercept equalled IN=8.91 dB and IN=10.92 dB for female and male voices, respectively.)

The effect of intentional raising of F0 on vocal intensity can be quantitatively analysed from the SPL-ESP-graphs as follows. Let us assume as depicted in Fig. 1 that the voice sample to be analysed is denoted as Sample1. The co-ordinates of this sample in the SPL-ESP-space are marked by SPL1 and ESP1, for SPL and ESP, respectively, and the equation for $\text{ESPlin}$ is given by $\text{ESPlin}=1.0\cdot\text{SPL}+\text{IN}$. The effect of F0 on the vocal intensity of Sample1 is now obtained as the (horizontal) difference between SPL1 and that point (denoted by an open circle) on $\text{ESPlin}$ whose ESP-value equals ESP1. This difference, denoted as $\text{SPL}_F0$, is obtained as a difference between SPL1 and ESP1 for SPL and ESP, respectively. The effect of fundamental frequency on the vocal intensity of a speech sample. ESPlin is a line in the SPL-ESP-space with a slope equal to 1. The black circles mark speech samples produced with an increasing intensity by a female speaker. The sample to be analysed is denoted as Sample1. The ESP-co-ordinate and ESP co-ordinate of this sample are denoted in the SPL-ESP-space as SPL1, and ESP1, respectively. The effect of fundamental frequency on the intensity of Sample1, denoted as $\text{SPL}_{F0}$, is obtained as a difference between SPL1 and that point (denoted by an open circle) on $\text{ESPlin}$ whose ESP-value equals ESP1.

### Table 1: Mean (m) and standard deviation (sd) for SPL$_{F0}$

<table>
<thead>
<tr>
<th>Gender</th>
<th>Phonation</th>
<th>m (dB)</th>
<th>sd (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>soft</td>
<td>0.0</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>normal</td>
<td>0.30</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>loud</td>
<td>4.08</td>
<td>2.88</td>
</tr>
<tr>
<td>Males</td>
<td>soft</td>
<td>-0.28</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>normal</td>
<td>0.07</td>
<td>1.76</td>
</tr>
<tr>
<td></td>
<td>loud</td>
<td>3.78</td>
<td>3.48</td>
</tr>
</tbody>
</table>

There are two possible mechanisms that explain the increase of vocal intensity caused by the raising of fundamental frequency. Firstly, increasing F0 implies that the number of instances of glottal closure per time unit increases. The fact that the closure of the glottis serves as the time instant of the main excitation of the vocal tract [9] implies that decay of the speech pressure waveform between any two instances of glottal closure will be less if F0 is increased. In other words, the raising of F0 causes the speech pressure waveform to contain more high-amplitude fluctuation, which increases SPL. (This phenomenon was called "Pulse repetition" by Fant [11].) Secondly, it is possible that increasing F0 corresponds to the speaker's tendency to take advantage of formant tuning [1], that is to say, the speaker adjusts F0 (or its harmonic) to coincide with a formant (typically the first formant) of the vocal tract. In order to find out which of these two techniques...
was used by the present speakers, we extracted from each speech sample the centre frequency of the first formant (F1). The frequency of F1 was obtained by looking for the first local resonance (in the vicinity of 700 Hz) from the all-pole spectrum of the digital vocal tract filter yielded by inverse filtering. The possible occurrence of formant tuning was analysed by computing, firstly, the difference between the first formant and the fundamental (d\(_{F0}=F1-F0\)) and, secondly, the difference between the first formant and the second harmonic (d\(_{2F0}=F1-2F0\)). If intensity is raised by adjusting the fundamental or its second harmonic to coincide with the first formant, either d\(_{F0}\) or d\(_{2F0}\) should approach zero. Similarly to [17], coincidence of F1 with a harmonic was assumed to occur if their difference was less than 50 Hz, which is a typical value for a formant bandwidth. The results obtained indicate that there was no systematic tuning of the harmonics to coincide with F1 when phonation was changed from soft to loud: all the speech samples yielded d\(_{F0}\)-values larger than 50 Hz, and only one speech sample produced in loud phonation by a female subject and two samples produced in loud phonation by males yielded d\(_{2F0}\)-values lower than 50 Hz. Hence, the increase of SPL caused by the effect of F0 cannot be explained by formant tuning. Instead, speakers used the increase of F0 as a method to amplify the high-energy fluctuations of the speech pressure waveform by increasing the number of glottal closures per time unit.

4. Conclusions

The speech material analysed in the present study included SPL-values over a large dynamic range of almost 60 dB. In changing SPL over such a large range, the subjects used surprisingly similar strategies in their intensity regulation. When intensity was increased from soft to normal, there was a moderate increase in the value of F0. For these voice samples, ESP came very close to being a linear function of SPL (with the slope of the line equal to unity). Hence, we can conclude that speakers tend to control their vocal intensity in soft and normal phonation by controlling mainly the shape and amplitude of glottal flow. F0 had only a minor effect on vocal intensity in these voice samples. This result, which was obtained in a quantitative form by ESP, is in line with a study by Gramming et al. [12], who considered the increase of F0 a passive consequence of an increase of sub-glottal pressure to modify the characteristics of the glottal flow.

In producing loud voice samples, the subjects changed radically their means of intensity regulation. This was seen especially in F0, whose value increased considerably. This raising of fundamental frequency resulted in a deviation of ESP-values from the linear function, which was followed by ESP (as a function of SPL) for voice samples in soft and normal phonation. This implies that the speakers used F0 in an active manner to increase vocal intensity. Hence, we may conclude that intentional raising of F0 is a means of intensity regulation used by speakers when loud voices need to be amplified. In the production of speaking voices, subjects do not take advantage of formant tuning in this procedure, but rather use F0 as a resource which helps to generate more high-amplitude fluctuations into the acoustic speech pressure waveform to increase vocal intensity.

5. References