LINEARITY OF THE FUNCTION BETWEEN THE SOUND PRESSURE LEVEL OF SPEECH AND THE NEGATIVE PEAK AMPLITUDE OF THE DIFFERENTIATED GLOTTAL FLOW FOR VOICES OF DIFFERENT INTENSITIES

Paavo Alku*, Juha Vintturri† and Erkki Vilkman†

*Helsinki University of Technology, Espoo, Finland
†Helsinki University Central Hospital, Helsinki, Finland

ABSTRACT

The negative peak amplitude of the differentiated glottal flow (d peak) is known to correlate strongly with the sound pressure level (SPL) of speech. Therefore, the function between d peak and SPL is usually modeled as a single line. In this survey, the relationship between d peak and SPL is revisited by analyzing glottal flows inverse filtered from speech sounds of largely different intensities. It is shown that in modeling SPL-d peak graphs the slope of the line matching soft phonation is larger than the slope of the line for normal and loud speech. This result suggests that vocal intensity is affected not only by the single amplitude domain value of the voice source, d peak, but also by the shape of the differentiated glottal flow near the instant of the negative peak.

1. INTRODUCTION

The role of the glottal source in regulating vocal intensity has been studied extensively during the past decades. The key factor behind intensity regulation of speech is subglottal pressure [4, 14]. Increasing intensity of speech implies raising subglottal pressure. However, also other factors of speech production, such as glottal adduction and formant frequencies, contribute to intensity regulation [12]. Increasing vocal intensity by raising subglottal pressure amplifies in general the AC-flow of a glottal pulse but it also makes the shape of the glottal pulse sharper so that it contains more energy at high frequencies. This implies that intensity of a soft sound is primarily determined by the fundamental whereas intensity of normal and loud speech is greatly affected by the overtones near the first formant [6, 8, 12]. Increasing subglottal pressure affects not only the shape of a single glottal pulse but also the repetition rate of the glottal pulses, i.e., the fundamental frequency (F0) of the voice. In [9] it was reported that the mean pitch increased by about half-semitones when intensity was increased by one decibel. According to [9] the increased value of F0 can be considered a passive result of raising subglottal pressure in order to produce louder sounds. However, F0 per se can also have an important role in intensity regulation of speech when formant tuning is used [11, 14, 15]. This implies that F0 of the voice is adjusted so that one of its lowest harmonics almost coincides with the first formant (F1). Consequently, the level of the speech spectrum at the first formant is amplified which increases the overall intensity. In order to increase vocal intensity with formant tuning calls for using large values of F0. Hence, formant tuning is a method of intensity regulation that is used mainly in singing or in production of high-pitched voices.

One of the most important parameters of the glottal flow, closely related to vocal intensity, is the negative peak amplitude of the differentiated glottal flow (d peak in Fig. 1). (This parameter is also called the maximum flow declination rate [10]). In [7] it was shown that there is a strong linear correlation between d peak (when expressed in dB units) and SPL (also expressed in dB units). The linearity of the function between SPL and d peak is further studied in the present survey. Even though there are well-known publications on this issue [e.g., 7], we consider this topic worth revisiting due to the following two rationales. First, it is known that dynamics of the human voice in terms of SPL can be 70 dB or even more in the case when pitch is allowed to vary over its extreme values [1, 14]. However, very soft or extremely loud voices are seldom analyzed in inverse filtering studies. Therefore, we believe that a further study on the SPL-d peak-function is needed in order to determine whether the linearity of the function holds when voices of largely different SPL-values are analyzed. The second rationale for the present study, as shown by Fig. 2 and Fig. 4, is our experimental finding indicating a clear “knee” in the SPL-d peak-graphs when voices are analyzed using a wide SPL-range. According to this finding SPL-d peak-graphs can be modeled accurately by a linear function but the slope of the line is different between soft and loud phonations.

2. MATERIAL AND METHODS

2.1. Speech material

In order to measure the linearity of the SPL-d peak-function, we designed an experiment where speech data were collected from eleven adult Finnish speakers (five females and six males). Each subject produced a series of the word /pa:p:a/ (Finnish for “grandpa”) by gradually increasing loudness. The acoustic speech pressure waveform was recorded using a condenser microphone (Brüel&Kjær 4176) which was placed at a distance of 40 cm from the lips of the speaker. Recordings were made in an anechoic chamber. 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&Kjaer 2225). By following the LED light display of the sound level meter, the speakers were able to control the SPL-values of their speech samples. Subjects repeated /pa:p:a/-words by increasing the SPL-values in
gradations of approximately 5 dB from the softest voice up to the loudest, with an SPL-value of 105 dB. The subjects were given no other restrictions regarding their voice production, which means that pitch and phonation type was decided freely by the speakers during the recording. The average number of voice samples covering the intensity range from the softest to the loudest phonation was 12 per speaker. The total number of speech samples produced was 61 and 71 by female and male speakers, respectively.

The acoustical speech pressure waveform was digitized using a sampling frequency of 22050 Hz and a resolution of 16 bits. At the computer, the signals were first high-pass filtered (cut-off frequency of the filter equalled 60 Hz) in order to remove any possible low frequency air pressure variations picked up during the recordings. The bandwidth of signals was 11 kHz. SPL-values were computed on the dB-scale for all speech signals using the RMS-operation (root mean square) and the SPL-value of the calibration tone (94 dB).

2.2. Inverse filtering
In order to estimate the glottal volume velocity waveforms, we used an inverse filtering technique that estimates the glottal excitation directly from the acoustic speech pressure waveform (i.e., no flow mask is required) [2]. Modeling of the vocal tract transfer function was done in inverse filtering using a sophisticated all-pole modeling technique, Discrete All-Pole Modeling (DAP) [5] instead of the conventional Linear Predictive Coding (LPC). The formants of the vocal tract, particularly the first formant, can be more accurately estimated by DAP than by LPC, which decreases the amount of formant ripple in the estimated glottal flows [2]. By scaling the DC-gain of the digital filter that models the vocal tract to unity, it is possible to estimate the amplitude characteristics of the glottal flow (with arbitrary units), even though no flow mask is used [3]. From each of the estimated glottal waveforms, the value of $d_{\text{peak}}$ was determined by computing the mean of negative peak amplitudes of the differentiated glottal flow over four consecutive glottal cycles.

### 3. RESULTS

An example of an $\text{SPL-}d_{\text{peak}}$-graph obtained from phonations of a male subject is shown in Fig. 2. At first sight, the function between $\text{SPL}$ and $d_{\text{peak}}$ seems to be linear over the whole $\text{SPL}$-range for the softest phonation up to the loudest. However, a closer examination of the graph reveals that $d_{\text{peak}}$ seems to follow very closely a linear function of $\text{SPL}$ over the softest four phonations but then, in the vicinity of 80 dB, it starts to follow another linear function, the slope of which is clearly smaller than the slope of the first line. In other words, if voices of greatly different SPL-values are analyzed, it seems to be more plausible to model the $\text{SPL-}d_{\text{peak}}$-relationship using (at least) two linear functions instead of just one.

The reason for the “knee” in the $\text{SPL-}d_{\text{peak}}$-graph between soft and normal phonations as shown by the example depicted in Fig. 2 can be explained by analyzing the waveforms of the corresponding glottal flows and their derivatives. The derivative of the glottal flow is shown for speech sample no. 4 of Fig. 2 (i.e., the strongest of the soft phonations) in Fig. 3(a). The differentiated glottal flow of speech sample no. 5 of Fig. 2 (i.e., the softest of the normal phonations) is shown by Fig. 3(b). It can be noticed from these graphs that the amplitude of $d_{\text{peak}}$ increases to some extent (by 25 % which equals 1.9 dB) when intensity of the voice rises. However, there is also a change in the shapes of the differentiated glottal flows: the waveform of Fig. 3(a) is much smoother during the glottal closing phase than in the signal shown in Fig. 3(b). In the frequency domain, this implies that the spectral decay of the differentiated glottal flow shown in Fig. 3(b) is less than the decay of the signal in Fig. 3(a). Therefore, the waveform of Fig. 3(b) produces, after being filtered through the vocal tract, a voice signal of larger $\text{SPL}$ than the waveform shown in Fig. 3(a). This is explained by the fact that the amplitude of the spectral components near the first formant of the produced speech sound will by stronger if the spectral decay of the voice source decreases. In this example a large increase of $\text{SPL}$ occurs even though the glottal flows have only a slightly different values of $d_{\text{peak}}$. When these voices are expressed in the $\text{SPL-}d_{\text{peak}}$-graph, they will have almost the same value on the y-axis (i.e., the level of $d_{\text{peak}}$), but the sound excited...
by the waveform of Fig. 3(b) will yield a larger value on the x-axis (i.e., the value of SPL). Hence, the accuracy of the classical model for the SPL-dpeak-function that consists of a single line is deteriorated. Therefore, it is justified to analyze how to model the SPL-dpeak-relationship more accurately by taking into account that, particularly in the soft-to-normal transitions, the change of the SPL-value of speech is regulated not only by the level of the negative peak of the differentiated glottal flow, dpeak, but also by the shape of the differentiated glottal waveform near the instant of the negative peak.

Figure 3. Glottal flow derivatives for speech sample no. 4 (graph a) and speech sample no. 5 (graph b) of Fig. 2.

SPL-dpeak-graphs of each speaker were modeled using two optimal linear functions. These lines, denoted by lineopt,1 and lineopt,2, have been drawn in an example of an SPL-dpeak-graph shown in Fig. 4. The functions of the optimal lines were determined for phonations of each speaker as follows: The obtained 12 SPL-dpeak-values of each subject were divided into two groups, denoted by Group1 and Group2. Group1 consisted firstly the SPL-dpeak-values of the three softest phonations while the rest of the nine SPL-dpeak-values were in Group2. An optimal linear function (in terms of the mean square error criterion) was then matched separately over the SPL-dpeak-values of both of the groups. Next, the obtained data points were divided into the groups differently by taking the four softest phonations into Group1 and the eight loudest phonations into Group2. A new pair of optimal linear functions was obtained by separately matching the SPL-dpeak-values of both groups with two lines. This procedure was repeated by searching for the division of the data points into two groups yielding a minimal mean square error between the original SPL-dpeak-values and their linear models. Hence, lineopt,1 and lineopt,2 shown in Fig. 4 represent linear models for the SPL-dpeak-values, including the optimal way to divide the data points into two separate groups to be matched by two lines.

The obtained SPL-dpeak-values of all the twelve subjects were analyzed in a similar manner as shown in Fig. 4. The optimal linear functions, lineopt,1 and lineopt,2, were quantified using their slopes which are denoted by slope1 and slope2, respectively. It was found that slope1 was larger than slope2 for the phonations of each subject. In other words, when SPL-dpeak-graphs were modeled by two linear functions that were determined optimally by minimizing the mean square error, the linear model for the softest phonations was a more steeply ascending line than the model matching normal and loud phonations in all cases. The mean (m) and the standard deviation (sd) of the two slopes were for female voices as follows: slope1: m = 0.82, sd = 0.29, slope2: m = 0.36, sd = 0.27. For male phonations, the following values were obtained: slope1: m = 1.14, sd = 0.27, slope2: m = 0.54, sd = 0.34. The difference between slope1 and slope2 was statistically tested using the Wilcoxon Signed-Rank nonparametric test. The difference between the slope values was statistically significant at the significance level of p=0.0033. The optimal linear functions modeled the SPL-dpeak-graphs accurately: the correlation coefficient averaged over the eleven subjects was 0.94 and 0.93 when SPL-dpeak-graphs were modeled with lineopt,1 and lineopt,2, respectively.

Figure 4. SPL-dpeak-graph, female speaker.

All the results reported so far in the present study are based on the glottal flow waveforms estimated by inverse filtering. Therefore, in order to confirm our results an additional analysis was made using the spectra of the radiated speech sounds per se (i.e., results given by inverse filtering were not used). By doing this frequency domain analysis we were able to compare our data with the previous results from phonetogram measurements [e.g., 8, 13]. We were interested in analyzing from the radiated spectra, whether the "knee" in the SPL-dpeak-function occurs simultaneously when the strongest partial changes from F0 to an overtone near F1. The following three speech samples of each subject were analyzed: the softest sound (denoted by s0(n)), the speech sample that occurs before the "knee" in the SPL-dpeak-function (denoted by s1(n)) and the sample that occurs straight after the "knee" (denoted by s2(n)). Spectrum was computed using FFT of 2048 samples with Hamming-windowing. From the spectra obtained it could be observed that the effect of the fundamental was largest for the intensity of the softest sound. However, the spectra of both s1(n) and s2(n) were characterized by strong
In order to quantitatively compare the effect of the fundamental on vocal intensity for $s_1(n)$, $s_2(n)$ and $s_0(n)$ we computed for each of these sounds the difference (in dB) between the overall energy and the energy without the fundamental. This difference was clearly largest for $s_0(n)$ (mean: 8.51 dB, standard deviation: 2.91 dB) when voices of all the subjects were analyzed. Both $s_1(n)$ and $s_2(n)$ yielded a value of energy difference that was much smaller and the value obtained for $s_1(n)$ (mean: 0.92 dB, standard deviation: 0.94 dB) was close to that computed from $s_2(n)$ (mean: 0.36 dB, standard deviation: 0.32 dB). This confirms that SPL reflects the amplitude of the strongest partial near F1 in the pitch-asynchronously computed spectra of $s_1(n)$ and $s_2(n)$. The value obtained was then compared to the frequency of the fundamental only for $s_0(n)$. However, SPL of both $s_1(n)$ and $s_2(n)$ is strongly affected by overtones near F1.

Finally, radiated speech spectra were also analyzed in order to test whether SPL of $s_2(n)$ was increased by formant tuning (i.e., by adjusting a harmonic closer to F1 in $s_2(n)$ than in $s_1(n)$). For this purpose we extracted the center frequency of F1 from the pitch-synchronously computed spectra of $s_1(n)$ and $s_2(n)$. The value obtained was then compared to the frequency of the strongest partial near F1 in the pitch-asynchronously computed spectra. This comparison showed that a spectral partial was closer to F1 in $s_1(n)$ in nine of the eleven cases, whereas an overtone was closer to F1 in $s_2(n)$ in only two of the eleven cases. This finding confirms our previous result according to which the "knee" in the SPL-$d_{peak}$-graph does not result from increasing intensity by formant tuning.

### 4. SUMMARY AND CONCLUSIONS

In previous studies, in particular in [7], it has been shown that the sound pressure level of speech follows the negative peak amplitude of the differentiated glottal flow, $d_{peak}$, in a manner close to linear. The linearity between SPL and $d_{peak}$ is re-addressed in the present study because of our finding indicating a clear "knee" in SPL-$d_{peak}$-graphs when voices of greatly different intensities are analyzed. This phenomenon was quantified in the present study by modeling the SPL-$d_{peak}$-graphs of eleven speakers with two optimal lines that minimize the mean square error between the overall energy and the energy without the fundamental only for $s_0(n)$. However, SPL of both $s_1(n)$ and $s_2(n)$ is strongly affected by overtones near F1.

In production of very soft voices, a speaker typically uses a smooth glottal pulse with a small AC-amplitude. Raising intensity can be achieved not only by increasing the AC-amplitude and shortening the closing phase of the glottal cycle [12]. Both of these changes in the glottal pulse increase the amplitude of the differentiated glottal flow. Results of the present study show that when intensity of speech is increased using minor SPL-steps, it is possible to generate two sounds with different SPL-values using practically the same level of $d_{peak}$, but the shape of the differentiated glottal flow is affected during the glottal closing phase. Hence, rising SPL can be achieved not only by increasing the amplitude of $d_{peak}$, as suggested by the classical SPL-$d_{peak}$-function reported in [7], but also by decreasing the spectral decay of the differentiated glottal flow by increasing the sharpness of the glottal flow derivative around the time-instant of $d_{peak}$. When voice intensity is changed from very soft to loud, speakers tend to make the most distinct change in their SPL-$d_{peak}$-function when going from "loud soft" to "soft normal". This change can be seen as a decrease in the slope of the line that matches the SPL-$d_{peak}$-graph.

### REFERENCES