ANALYZING THE EFFECT OF SECONDARY EXCITATIONS OF THE VOCAL TRACT ON VOCAL INTENSITY IN DIFFERENT LOUDNESS CONDITIONS

Paavo Alku[1], Juha Vintturi[2], Erkki Vilkman[3]
1: University of Turku, Vesilinnantie 5, FIN-20014 Turku, Finland
2: Helsinki University Central Hospital, Haartmanink. 4 E, Fin-00290 Helsinki, Finland
3: University of Oulu, Fin-90220 Oulu, Finland
e-mail: Paavo.Alku@hut.fi

ABSTRACT

For voiced speech the main excitation of the vocal tract occurs at the end of the glottal closing phase when the rate of change of the flow reaches its absolute maximum. This study presents a straightforward method that yields a numerical value to characterize the effect of the main excitation on vocal intensity. The method, Energy Ratio by Modified Excitation (*ERME*), takes advantage of the glottal flow and the model of the vocal tract transfer function given by inverse filtering and it synthesizes two signals based on the source-filter theory. The first synthesized sound is produced using the glottal flow waveform given by inverse filtering *per se*. The second signal is synthesized by removing the main excitation from the differentiated glottal flow. *ERME* is defined as the ratio between the energy of the first synthesized signal and the energy of the second one. It is shown that when the loudness of speech increases, the value of *ERME* first rises but in the case of loud voices it starts to decrease. This behavior of *ERME* shows that effects of secondary excitations of the vocal tract that occur during glottal opening become important in the production of loud voices.

1. INTRODUCTION

Gauffin and Sundberg [9] showed that the negative peak amplitude of the differentiated glottal flow (d_{peak}.) has a strong linear correlation with the sound pressure level of speech (*SPL)*. Correlation between d_{peak} and *SPL* can be explained with the help of the linear source-filter theory of speech production [5] as follows. According to [5] production of a voiced speech signal can be modeled by three separate processes: the glottal flow, the vocal tract filtering and the lip radiation effect. The lip radiation effect can be estimated as a differentiator. By combining the glottal flow and the lip radiation effect it is possible to model voice production by two processes: the differentiated glottal flow (also called the effective driving function) and the vocal tract filtering [e.g., 18]. A negative impulse like peak dominates the waveform of the differentiated glottal flow at least for normal and loud voices. This peak which occurs at the same instant when the rate of the change of the flow reaches its absolute maximum serves as the main excitation of the vocal tract [7]. Consequently, the level of the peak, d_{peak}, determines to a large extent the energy of the produced speech signal. This results in a strong correlation between d_{peak}, and the *SPL* of the produced speech signal. Behavior of the negative peak amplitude of the differentiated glottal flow in different loudness conditions has been studied widely, e.g. [7]-[11], [16], [17]. However, there are only a few studies where secondary excitations of the vocal tract has been discussed [3], [6], [12], [13]. The background for the present study is our finding according to which correlation between *SPL* and d_{peak} tends to decrease when loud voices are compared to those of normal loudness. This means that when d_{peak} (in dB units) is plotted as a function of *SPL* (in dB units) an almost linear function is obtained between the two quantities for speech samples of normal loudness. However, when loudness is increased d_{peak} seems to reach a saturation point after which it does not change considerably even though *SPL* grows. This saturation of d_{peak} has been reported by the present authors in [2] and also in [8]. We believe that the saturation of d_{peak} is explained by the fact that the parameterization of voice production with the negative peak amplitude of the differentiated glottal flow takes into account only the main excitation of the vocal tract. Hence, when applying d_{peak} the secondary excitations of the vocal tract that occur, for example, during the glottal opening have been completely ignored.

The rationale for this study is to present a method that makes possible measuring the effect of the main excitation on vocal intensity when loudness is changed from soft to loud. The emphasis is put on developing a method with which it would be possible to analyze to what extent vocal intensity is affected by secondary excitations especially for loud voices.

2. MATERIAL AND METHODS

2.1 Speech material and inverse filtering

An experiment was arranged where speech data were collected from five female and six male subjects. Each subject produced a series of the word /pa:p:a/ by gradually increasing loudness. The first phonation sample was produced as softly as possible without whispering. 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 the *SPL*-value of 105 dB. (Some subjects voluntarily also produced the loudest sound with the *SPL*-value of 110 dB.) Pitch and phonation type was decided freely by the speakers during the recording. In order to gather enough data representing great vocal intensity the subjects repeated the three loudest speech samples three times. The total number of speech samples produced was 86 and 102 by female and male speakers, respectively.

Two signals, the acoustic speech pressure waveform and the intraoral pressure, were simultaneously recorded in an anechoic chamber. The former was measured using a condenser microphone (Brüel&Kjaer 4176) that was placed at a distance of 40 cm from the lips of the speaker. The intraoral pressure was recorded using a pressure transducer (Frøkjær-Jensen Electronics, Manophone, MF710) that was connected to the mouth of the subject with a plastic catheter. Both the acoustical speech waveform and the pressure signal was digitized using the sampling frequency of 22050 Hz and the resolution of 16 bits. *SPL*-values were computed on the dB-scale (flat weighting) for all speech signals using the root mean square operation and the *SPL*-value of the calibration tone (94 dB). Estimates for sub-glottal pressure were computed from the oral-pressure waveforms during the p-occlusion using an approach similar to that presented in [15].

Speech signals were inverse filtered in order to obtain estimates for the glottal volume velocity waveforms. Estimation of the glottal source was computed directly from the acoustic speech pressure waveform (i.e., no flow mask was used) by applying a method that is similar to the one described in [1]. In the present study the estimation of the vocal tract transfer function was based on a sophisticated all-pole modeling technique, called Discrete All-pole Modeling (DAP) [4] instead of the conventional LP-analysis that is used in [1]. The inverse filtering method used estimates the vocal tract transfer function using the DAP-technique by first canceling the average effect of the glottal source from the speech spectrum using a low order all-pole filtering. The lip radiation effect was canceled by a first order all-pole filter with its pole in the z-domain at z=0.98. The estimation of the glottal flow was computed using an analysis window of 50 ms. The amplitude values of the estimated glottal flows were scaled using a method described in [2] that is based on setting the DC-gain of the vocal tract model in inverse filtering to unity

2.2 Parameterization of the effect of the main excitation on vocal intensity

The goal of the current study is to present a new straightforward method to quantify the effect of the main excitation (i.e., the dominating negative peak of the differentiated glottal flow that occurs during the glottal closing phase) on vocal intensity. The method takes advantage of the glottal flow and the model of the vocal tract transfer function that are estimated by inverse filtering. The source-filter theory makes possible modeling voice production with the following two processes: the differentiated flow serves as the excitation, and the digital all-pole model of the vocal tract serves as the filter.

Modeling voice production with the effective driving function and the vocal tract filter makes synthesizing speech signals possible. Our method is based on the idea of comparing, in terms of energy, two synthesized speech sounds using this source-filter model. In the first case the effective driving function corresponds to the derivative of the glottal flow given by inverse filtering and the filter is the all-pole model also given by inverse filtering. The synthesized speech signal, denoted $s_1(n)$, is obviously the same as the original signal from which the inverse filtering analysis was computed. In the second case the same model for the vocal tract filter is used but the effective driving function is modified by assigning the value of zero for all its negative data samples. This approach makes possible synthesizing a hypothetical speech signal, denoted by $s_2(n)$, that has been produced without having the dominating main excitation in the input waveform. The effect of resetting the main excitations from the effective driving function can be quantified by taking the ratio of the energy of signals $s_1(n)$ and $s_2(n)$. This energy ratio, Energy Ratio by Modified Excitation (*ERME*), is defined as follows:

$$ERME = \frac{\sum_{n=0}^{N-1} s_1^2(n)}{\sum_{n=0}^{N-1} s_2^2(n)} \qquad (1)$$

where N denotes the length of the analysis window

In the case of a speech sound the energy of which has been produced mostly by main excitations it is expected that resetting the negative peaks of the effective driving function greatly decreases the energy of the synthesized signal. In other words, the more important the role of the main excitation is in the production of vocal intensity the larger the value of *ERME* is expected to be. On the other hand a small value of *ERME* indicated that removing the main excitations from the effective driving function did not greatly affect the energy of the synthesized speech sound. Hence, in this case the original glottal excitation must have had important secondary excitations outside the glottal closing phase. It is worth noting that removing of the negative peak of the differentiated glottal flow creates a DC-component in $s_2(n)$ that has to be removed.

3. RESULTS

A graph depicting typical behavior of *ERME* as a function of *SPL* is shown in Fig. 1(a) for the voices of one male speaker. Estimated values for the sub-glottal pressure are presented in Fig. 1(b). Fig. 1(a) shows that the value of *ERME* is smallest for soft voices. This implies that the energy of the hypothetical signal $s_2(n)$ has not decreased very much after removing the peaks of the main excitations from the effective driving function. In other words, the role of the main excitation has not dominated in the production of intensity for the softest sounds because there has been an almost equal excitation (although opposite in sign) during glottal opening. This means that the shape of the glottal flow pulse is smooth in the case of soft voices, as reported in previous studies [e.g., 11], and the steepness of the closing phase is not very much larger than the steepness of the opening phase. As can be seen from Fig. 1(a) the value of *ERME* rises when *SPL* of speech is increased. The maximum of *ERME* occurs at the *SPL*-value of approximately 85 dB. This implies that in producing normal or slightly loud voices the effect of the main excitation on vocal intensity is largest. However, an interesting phenomenon occurs when *SPL* is further increased; the value of *ERME* starts to decrease. The reason for this is the steepening of the glottal opening phase due to the increased subglottal pressure. Hence, for the loudest voices the vocal tract gets a relatively strong excitation also at glottal opening. The decrease of *ERME* after *SPL*-values of approximately 85 dB occurred for each of the eleven analyzed subjects.

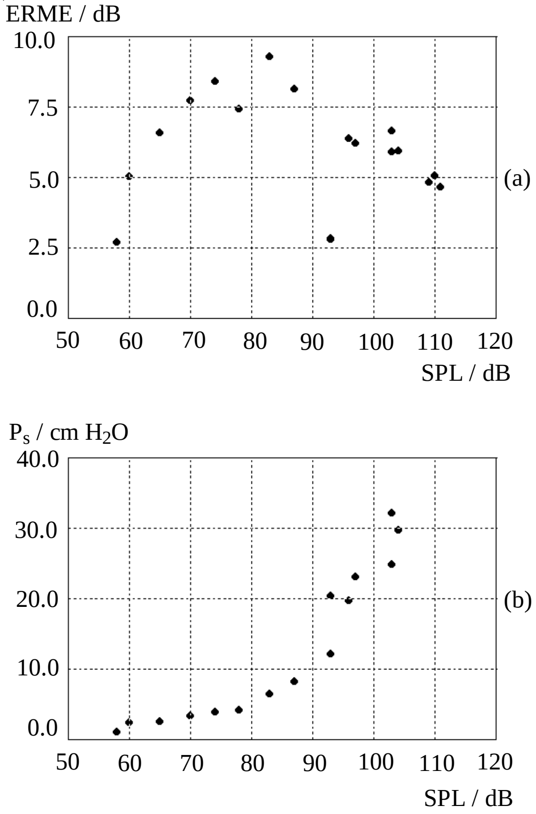

Figure 1: ERME (a) and sub-glottal pressure (b) as a function of SPL for a male speaker.

ERME-values are shown as a function of *SPL* for all the female and male speakers in Fig. 2. As noted above, the behavior of *ERME* as a function of *SPL* consists of an increasing part and a decreasing portion. Therefore, two linear functions were matched over the *ERME*-values computed from voices of the female and male speakers. (The optimization included dividing the data into two groups: the first group that was modeled by an ascending regression line and the second group that was modeled by a descending line. The optimal way to divide the data points into two groups was determined by searching for the division that minimizes the mean square error between the data points and the two regression lines.) Correlation coefficients (*r*) for the ascending regression lines were 0.60 for the female speakers and 0.81 for the males. For the descending regression lines correlation was clearly smaller: *r*= -0.27 for females and *r*=-0.30 for males.

Correlation between *ERME* and *SPL* was also studied by analyzing voices of each individual subject separately. For the ascending part of the data the obtained *r*-values varied between 0.85 and 0.95 for female speakers and the mean *r*-value was 0.91. For male speakers the correlation coefficient was between 0.84 and 0.97 and the mean *r*-value equaled 0.91. When correlation was analyzed for female subjects over the descending part of the data the *r*-values varied between -0.10 and -0.55 and the mean *r*-value equaled -0.23. For male speakers the *r*-value of the descending part of the data varied between -0.12 and -0.82 and the mean was -0.39.

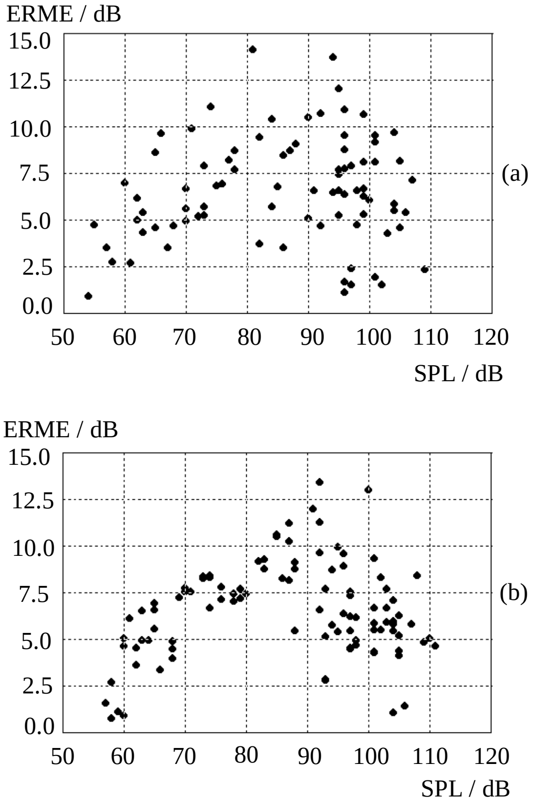

Figure 2: ERME as a function of SPL for all the female (a) and male (b) speakers.

4. SUMMARY AND CONCLUSIONS

Inverse filtering yields estimates for both the glottal flow and the effect of the vocal tract filtering. Therefore it is possible to synthesize the original speech signal by filtering the differentiated flow (the effective driving function) with the digital filter that models the vocal tract transfer function. Using this speech production model it is also possible to synthesize hypothetical speech waveforms by manipulating the effective driving function. The aim of this study was to analyze the effect of the main excitation of the vocal tract on vocal intensity. Therefore, the model mentioned above was used to synthesize hypothetical speech signals by using an input where the main excitation of the vocal tract is removed by resetting the negative samples of the effective driving function. A numerical value, *ERME*, for the effect of the main excitation on vocal intensity was obtained by computing the energy ratio between the original speech signal and the hypothetical one.

Speech material analyzed included *SPL*-values over a large dynamic range: the minimum *SPL* was 54 dB and the maximum *SPL* was 111 dB. The main finding of the study was that *ERME* showed a similar trend for the voices of all the speakers: the value of *ERME* first increased but then, in the vicinity of 85 dB, it started to decrease rapidly as a function of *SPL*. During the ascending portion there was a strong linear correlation between *ERME* (in dB units) and *SPL* (in dB units) for all the analyzed eleven speakers.

The maximum of *ERME* occurred approximately at *SPL*-values of 85 dB. This implies that the effect of the main excitation on vocal intensity has been largest and secondary excitations have been of minor importance in power regulation of speech. This results from the rate of change in the glottal flow that is greater during the closing phase than during the opening one. On the other hand, the effect of the secondary excitation during glottal opening on vocal intensity is decreased due to losses of the vocal tract that increase during the open phase of the glottis [7], [14]. Consequently, the short-time energy of the speech signal produced during the glottal open phase will be smaller in comparison to the energy over the glottal closed phase that is characterized by strong formant oscillations due to the main excitation of the vocal tract.

The obtained values of *ERME* started to decrease rapidly after the *SPL*-value of approximately 85 dB for all the subjects when loudness was increased. This behavior of *ERME* is explained by the secondary excitations of the vocal tract that occur especially during the opening of the vocal folds. Production of speech with *SPL* over approximately 85 dB corresponded to a noticeable increase of the subglottal pressure. Consequently, the shape of the glottal pulse became steeper during the glottal opening phase and at the same time the time span of the glottal open phase decreased due to the increase of F0. The relative importance of the rapid changes in the flow during the opening phase increases and at the same time the damping effect of the vocal tract decreases due to the shortening of the time duration of the glottal open phase. It is the increase in the amplitude of these secondary excitations that caused the value of *ERME* to decrease at *SPL*-values over 85 dB. The increased importance of the secondary excitations was confirmed by comparing *ERME* to sub-glottal pressure values as a function of *SPL*. In this comparison it was found that the turning point in the *ERME*-graphs seemed to coincide with the *SPL*-values when the subjects started greatly to increase their subglottal pressure in order to produce voices of increased loudness.

This survey serves as a preliminary study in order to analyze the role of the secondary excitations of the vocal tract in intensity regulation of speech when *SPL* is changed over a wide dynamic range. From the results of the study it is obvious that the effects of the secondary excitations differ greatly when voices of normal loudness are compared to loud speech. Further studies with a larger number of speakers are needed in order to get more detailed information of, for example, the difference in the role of the secondary excitations between female and male speakers.

ACKNOWLEDGMENTS

This study has been supported by the Academy of Finland.

5. REFERENCES

1. Alku, P. “Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering,” Speech Comm. 11: 109-118, 1992.
2. Alku, P., Vilkman, E., and Laukkanen, A-M. “Parameterization of the voice source by combining spectral decay and amplitude features of the glottal flow,” J. Speech Lang. Hear. Res. (In press), 1998.
3. Childers, D., and Lee, C. “Vocal quality factors: Analysis, synthesis, and perception,” J. Acoust. Soc. Am. 90: 2394-2410, 1991.
4. El-Jaroudi, A., and Makhoul, J. “Discrete all-pole modeling,” IEEE Trans. Signal Proc. 39: 411-423, 1991.
5. Fant, G. *Acoustic Theory of Speech Production* (Mouton, the Hague), 1960.
6. Fant, G. “Glottal source and excitation analysis,” Speech Transmission Laboratory, Stockholm, Sweden: Royal Institute of Technology, QPSR 1:85-107, 1979.
7. Fant, G. “Some problems in voice source analysis,” Speech Comm. 13: 7-22, 1993.
8. Fant, G., Hertegård, S., Kruckenberg, A, and Liljencrants, J. “Covariation of subglottal pressure, F0 and glottal parameters,” Proc. 5th European Conf. on Speech Communication and Technology, 1: 453-456, 1997.
9. Gauffin, J., and Sundberg, J. “Spectral correlates of glottal voice source waveform characteristics,” J. Speech Hearing Res. 32: 556-565, 1989.
10. Hertegård, S., Gauffin, J., and Karlsson, I. “Physiological correlates of the inverse filtered flow waveforms,” J. Voice 6: 224-234, 1992.
11. Holmberg, E., Hillman, R., and Perkell, J. “Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice,” J. Acoust. Soc. Am. 84: 511-529, 1988.
12. Holmes, J. “Formant excitation before and after glottal closure,” Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., 1: 39-42, 1976.
13. Hunt, M. “Studies of glottal excitation using inverse filtering and an electroglottograph,” Proc. of the 11th Int. Congress on Phonetic Sciences, 3: 23-26,1987.
14. Klatt, D., and Klatt, L. “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am. 87: 820-857, 1990.
15. Löfqvist, A., Kitzing, P., and Carlborg, A. “Initial validation of an indirect measure of subglottal pressure during voicing,” J. Acoust. Soc. Am. 72: 633-635, 1982.
16. Sundberg, J., Titze, I., and Scherer, R. “Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source,” J. Voice 7: 15-29, 1993.
17. Titze, I., and Sundberg, J. “Vocal intensity in speakers and singers,” J. Acoust. Soc. Am. 91: 2936-2946, 1992.
18. Wong, D., Markel, J., and Gray, A. “Least squares glottal inverse filtering from the acoustic speech waveform,” IEEE Trans. Acoust. Speech Signal Proc. 27: 350-355, 1979.