6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

GLOTTAL PARAMETERS CONTRIBUTING TO THE PERCEPTION OF LOUD VOICES

*Sopae Yi, **Hyung Soon Kim, ***One Good Lee

*Department of Cognitive Science, **Department of Electronics Engineering,
***Department of English Literature and Language, Pusan National University
San 30 Jang-Jun Dong, Gum-Jung Goo, Pusan, 609-735, Korea

Email : spyi@web.pusan.ac.kr

10.21437/ICSLP.2000-878

ABSTRACT

This paper focused on glottal parameters contributing to the perception of loud voices because energy of a voice is not the only effective factor. We used a formant synthesizer to synthesize loud voices. We divided F0 tilt (the tilt of F0 contour), SQ (Speed Quotient), OQ (Open Quotient) and TL (spectral Tilt Level) into three levels to get different combinations with default values for the other synthesizer parameters. Analysis of listening tests indicates that F0 tilt, SQ, OQ and TL in descending order had significant influence on the perception of loud voices. F0 tilt had far more significant effect than the others. The influence of SQ increased a lot with exclusion of F0 tilt as a factor. The interaction between parameters was not significant.

1. INTRODUCTION

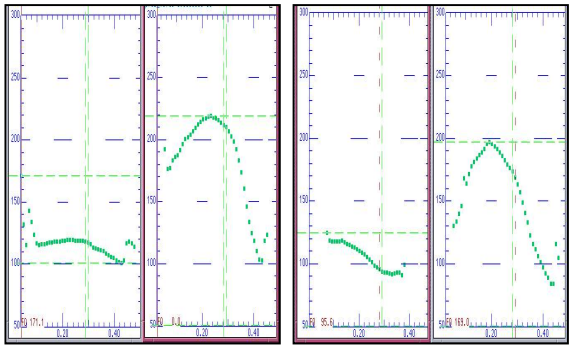
As intelligibility of a synthesized speech has improved, naturalness has become a more important issue. Many attempts are being made to improve the quality of synthetic voices. Efforts are being made even to reflect different emotions in synthesized speech. One of the ways to carry emotions in a speech can be adjusting the loudness of a voice. This paper focused on the perception of loud voices, one of the paralinguistic speech types, by experimentally studying the glottal parameters. In doing so, we also studied the way to improve the quality of synthesized speech by using a formant synthesizer.

We used the Klatt88 synthesizer with LF-model as its source [3] to study the contribution of glottal parameters to the perception of loud voices. We studied F0 tilt, SQ, OQ and TL. Loud voices in natural sound tend to have a shorter duration [1]. This paper, however, does not take this 'shortness of duration' phenomena into consideration.

We discuss acoustic characteristics of loud voices in section 2. We briefly describe our experimentation and analysis in section 3 and 4. We discuss the result in section 5.

2. ACOUSTIC CHARACTERISTICS OF LOUD VOICES

According to the analysis of natural voices, loud voices have higher F0 (Fundamental Frequency), SQ and lower OQ than normal voices [1]. Figure 1 shows F0 contours of normal voices and loud voices of /a/ vowels. These F0 contours are from two individuals out of the ten participants whose F0 values revealed similar patterns. Figure 1 shows the overall F0 values of loud voices higher and more dynamic than normal counterparts. F0 contours of loud voices rise more steeply in the beginning and fall more steeply in the end than the contours of normal voices.



Normal Loud Normal Loud

(a) Speaker KJH (b) Speaker KSJ

Figure 1: F0 contours of /a/ vowels in loud and normal voices. (Horizontal and vertical axes represent time in seconds and F0 values in Hz, respectively.)

Figure 2 shows a glottal flow signal, its derivative and LF parameters. U_0 is the maximum of Ug. E_e is the absolute value of the minimum of dUg. Glottis begins to open at time t = 0. Glottis begins to close at time t_c. The time points of U_0 and E_e are t_p and t_e respectively. The time period between t_e and the projection of the tangent of dUg at t_e is t_a. t_n is equal to $t_e - t_p$. OQ is the ratio of open time to total period duration, e.d. OQ = $(t_c - 0) / t_0$. SQ is the ratio of the duration of the rising portion to the duration of the falling portion of the glottal open phase, e.d. SQ = $t_p / (t_e - t_p)$. OQ influences the relative energy level of the first harmonic [2] while SQ is related to the energy level of first, second and third harmonics [5][6]. Increasing TL (spectral Tilt Level) attenuates high-frequency components associated with "corner rounding" resulting from the non simultaneous closure along the length of the vocal folds [2].

The glottal waveforms in loud voice show sharp angles between the end of the closing and the beginning of the closed portions [1] which result in the boost of the high frequency components

suggesting the decrease of TL [2]. The closed portion in a loud voice is often well defined and sometimes relatively longer, which would result in a smaller OQ [1]. SQ in a loud voice is greater than the one in a normal voice [1]. A loud voice was typically produced with higher fundamental frequency than a normal voice [1].

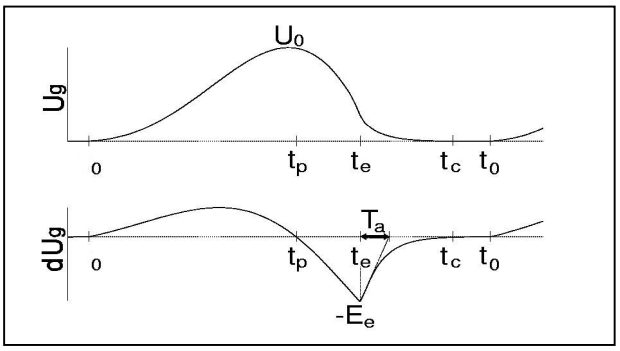


Figure 2: Glottal flow (Ug) and glottal flow derivative (dUg) with the parameters of LF-model [5].

3. EXPERIMENT

We used the Klatt88 synthesizer with LF-model as its source [3] to generate vowels. We divided F0 tilt, SQ, OQ and TL into three levels to get different combinations with default values for the other synthesizer parameters. We also normalized the energy of all stimuli to exclude energy as a perceptual factor. The F0 values are based on the data from natural voices.

We synthesized the /a/ vowels by using the average first, second and third formant values from 76 male speakers [8]. Table 1 shows F0 values in each level. Five values for each level indicates each vertex of piecewise linear curves (see Figure 3). Piecewise linear curves are used to approximate the original F0 curve. The piecewise linear curve and the original F0 curve showed almost no perceptual difference according to the preliminary listening test.

Table 1: Values of vertexes of piecewise linear F0 curve for each level

	(25 ms)	(155 ms)	(285 ms)	(500 ms)	(600 ms)
Level 1	136.2 Hz	146 Hz	156 Hz	146 Hz	102 Hz
Level 2	159.5 Hz	220.3 Hz	237.5 Hz	223 Hz	102 Hz
Level 3	182.8 Hz	294.6 Hz	319 Hz	300 Hz	102 Hz

To avoid the argument that the piecewise linear curves are types, not values, we computed F0 tilt as follows:

$$\text{F0 tilt (Hz/ms)} = (\text{Max\_F0} - \text{Initial\_F0}) / (T2 - T1) \qquad (1)$$

where Initial_F0 is the starting value of the F0 contour, and Max_F0 is the maximum value of the F0 contour, T1 and T2 are the time values of Initial_F0 and Max_F0 respectively (e.g. F0 tilt for level 1 = (156 Hz – 136.2 Hz) / (285 ms – 25 ms)). The equation (1) is based on the report that the dynamic component of F0 contour influences the perception of the loudness (7). The greater F0 tilt, the more dynamic the contour becomes.

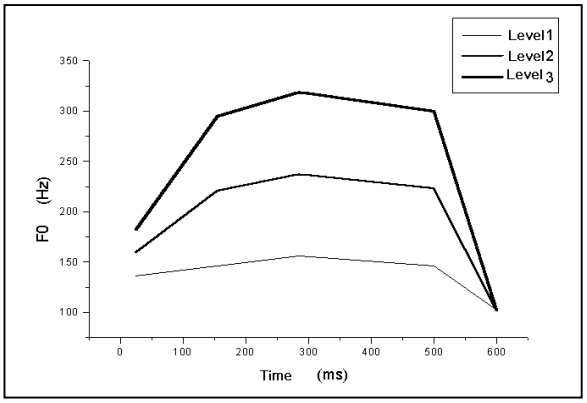


Figure 3: Three F0 curves used to synthesize /a/ vowels

The parameter values of each level are shown in table 2. OQ values and SQ values at level one and three are the minimum and maximum values of OQ and SQ of 25 male speakers' glottal waveform [1]. OQ values and SQ values at level two are averages of level one and level three. TL values are divided into 12dB, 6dB and 0dB [3].

Table 2: Parameter values at three levels

	F0 tilt	SQ	OQ	TL
Level 1	0.124	1.32	4.6	0 dB
Level 2	0.488	2.19	6.2	6 dB
Level 3	0.851	3.06	7.8	12 dB

F0 tilt, SQ values at level 1 and OQ, TL values at level 3 with the rest of the parameters at default values are used to make a normal sound [1]. We increased the AV (Amplitude of Voicing) of this sound from 60dB to 70dB (maximum AV is 80dB) to make a reference sound labeled A. In doing this, we intended to show that amplitude variation is not the only effective factor for the perception of loud voices. If amplitude variation is a dominantly contributing factor for loud voice perception, reference sound A which has increased amplitude will be enough for the loud voice perception.

Reference sound B is made up of F0 tilt, SQ at level 3 and OQ, TL at level 1. Reference sound B has the acoustic characteristics

of loud voices in natural sound [1]. According to Humbert et al., loud voices have smaller OQ, TL and larger F0, SQ values than normal voices [1]. Listeners participating in a preliminary listening test came to the unanimous consensus that the reference sound B is louder than the reference sound A. The rationale behind making reference sounds A and B is to avoid the subjects' biased judgement.

Figure 4 is the picture of the computer interface used for the listening test. The computer interface used in this study adopted an analogue scale method. Compared with a discrete scale method, an analogue scale method enhances the consistency of the listeners' judgement [4]. With this interface, the whole stimuli can be seen and compared with ease. Any stimulus can be found and played at any time .

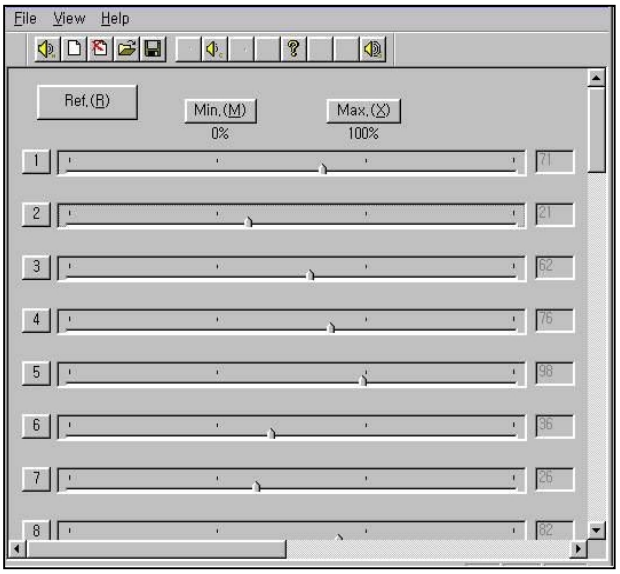


Figure 4: Computer interface used for the listening test

Figure 4 shows upper part of 81 slide bars, numbered from 1 to 81. On the right end of each bar is a box displaying the value which changes according to the position of the button on the bar. Listeners clicked the button with a number on the left of each bar and heard the /a/ sound assigned to it. On top of the bars are the buttons with Min (0%) and Max (100%) written on them. The button with Min (0%) and the button with Max (100%) are assigned to the reference sound A and the reference sound B respectively. The listeners compared 81 sounds with the reference sounds A and B. They also compared the 81 sounds with each other.

If a test sound is louder than the reference sound A, listeners dragged the button on the bar to the right of the button assigned to the reference sound A. If the reference sound A is louder than a test sound, listeners dragged the button of the test sound to the left of the reference sound A. The louder a sound is the farther the button should move to the right. The perceptual values getting out of the range from 0% to 100% are to be covered in the range from –100% to +200%. The four lines on each bar indicate positions for –100%, 0%, +100% and +200% from left to right.

For example, if a button assigned to a sound has the same perceptual loudness as the reference sound A, listeners are supposed to drag the button on the bar to the position of 0% where the 'Min' button is located. The position of each button is determined by the relative loudness in comparison with other sounds including the reference sounds A and B. The more similar the degree of perceptual loudness between sounds, the closer the distance between the buttons. Listeners clicked a button as many times as they want to listen to the sounds.

The first listening test was done with four parameters (F0 tilt, SQ, OQ and TL) of three levels making 81 combinations. The second test is done with three parameters (SQ, OQ and TL) of three levels making 27 combinations. Ten listeners without any hearing problem participated in the tests. All of them are in their twenties and not familiar with the synthesized sounds being evaluated. In this way, we hope to avoid any influence on the results based on previous experience. Sounds are sorted at random each time. Two sounds which have the same parameter combination as the reference sounds A and B are also included in the 81 sounds to be tested. Some of the data are excluded if a listener judges the two sounds as different from the reference sounds, since the validity of the judgement is questionable.

4. ANALYSIS OF THE EXPERIMENTS

4.1. First Experiment

The four parameters (F0 tilt, SQ, OQ, and TL) are used to make 81 combinations (3x3x3x3). F values and P values are obtained from the multi-way factorial design of the first experiment. The interaction between parameters is not statistically significant at 1%. F0 tilt, SQ, OQ and TL are statistically significant at 1%.

Table 3: Multiple regression analysis of the first experiment

	Multiple regression coeff.	**Determi-nation coeff.**	**F values**	**P values**
TL	0.29	0.0018	6.06	**0.0141**
OQ	-1.91	0.0055	18.93	**0.0001**
SQ	8.05	0.0289	97.24	**0.0001**
F0 Tilt	96.88	**0.7310**	2195.49	**0.0001**

We used partial correlation coefficients to judge the trend and linearity between the perception of loud voices and other parameters. The relationship between the perception of loud voices and the four parameters are statistically significant at 10%. F0 tilt has the highest linearity with the perception of loud voices. SQ has the second highest linearity, OQ the third and TL the fourth. F0 tilt, SQ and TL have a positive relationship

with the perception of loud voice and OQ has a negative relationship.

Multiple regression analysis reveals that all parameters are statistically significant at 10% (see Table 3). The contribution to the perception of loud voice by the parameters is estimated. According to the amount of contribution, F0 tilt (73%), SQ (2%), OQ (0.5%) and TL (0.1%) can be listed in descending order. F0 tilt seems to be the dominating contributor to the perception of loud voices. It is reasonable to say that listeners judged the loudness by using the acoustic cue of F0 tilt.

4.2. Second Experiment

We tried to find out the perceptual effect of the parameters without the influence of F0 tilt. We made 27 combinations (3x3x3) out of the three parameters, SQ, OQ and TL. F0 tilt was fixed at level 2 which means we used only one level of F0 tilt. As in the first experiment, the factorial design analysis and partial correlation analysis of the second experiment show a similar trend. The interaction between parameters is not statistically significant at 1%. SQ, OQ, TL are statistically significant at 1%. The partial correlation between SQ, OQ, TL and the perception of loud voice is statistically significant at 1%. SQ and TL have a positive relationship with the perception of loud voices whereas OQ has a negative relationship.

Table 4: Multiple regression analysis of the second experiment

	Multiple regression coeff.	**Determi-nation coeff.**	**F values**	**P values**
TL	1.89	0.0347	21.21	**0.0001**
OQ	-7.08	0.0346	22.88	**0.0001**
SQ	50.82	**0.5281**	299.91	**0.0001**

Multiple regression reveals that SQ, OQ and TL are statistically significant at 1% (see Table 4). The contribution of SQ becomes much greater than before because of the exclusion of F0 influence. We concluded that the dominant influence of F0 tilt mitigates the significant difference between the three parameters, SQ, OQ and TL in the first experiment and the acoustic cue of SQ played a major role in the listeners' judgement of loudness in the second experiment.

5. DISCUSSION

Comparing the sum of values at each level, we found that F0 tilt at level 3, SQ at level 3 (3.06), OQ at level 1 (0.46) and TL at level 2 (6dB) make the optimal combination contributing to the perception of loud voices. The dynamic characteristic of F0 tilt showed a dominant influence on the perception of loud voices. Neuhoff et al. reported that the dynamic characteristic of the F0 contour contributes to the perception of loudness [7]. The influence of SQ is dominant in the second experiment in which the influence of F0 tilt is excluded.

The increase of SQ is known to reduce the energy level of first, second and third harmonics leading to the vocal quality of pressed voices [5][6]. Therefore the increase of SQ contributed to the perception of loud voices by increasing the auditory effect of pressed voices. While F0 tilt has a direct influence on the loudness, SQ, OQ and TL seem to influence the vocal quality found in loud voices.

Rather than comparing various vowels only /a/ vowels are used in this paper since categorical influence of vowels has no significant impact on the loudness. The glottal parameters of loudness are mainly influenced by the glottal status, not by the kind of vowel. Glottal condition of a phonation, however, changes at word levels and sentence levels. Our next study should expand the range from a single vowel to words and sentences.

REFERENCES

1. E. B. Holmberg, R. E. Hillman and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements of male and female speakers in soft, normal and loud voice," *JASA*, vol. 84, no.2, pp.511-529, 1988.
2. D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *JASA*, vol.87, no.2, pp.820-857, 1990.
3. D. Klatt, "Description of the cascade/parallel formant synthesizer," *KLATTALK, Conversion of English Text to Speech*, Chapter 3, 1990.
4. S. Granqvist, "Enhancements to the Visual Analogue Scale, VAS, for listening tests," TMH-QPSR 4/1996, 1996.
5. H. Strik, *Physiological control and behavior of the voice source in the production of prosody*, Katholieke Universiteit Nijmegen, 1994.
6. I. Karlsson, "Voice source dynamics for female speakers," Proceedings of the 1990 International Conference on Spoken Language Processing, Kobe, pp.69-72, 1990.
7. J, Neuhoff and M. McBeath, "The Interaction of pitch and loudness in dynamic stimuli: beyond the doppler illusion,"133rd ASA Meeting, 1997.
8. G. E. Peterson and H. L. Barney, "Control methods used in a study of the identification of vowels," JASA, vol.24, pp.175-184, 1954.