The perceptual relevance of glottal-pulse parameter variations

Ralph van Dinther, Raymond N.J. Veldhuis and Armin Kohlrausch[1]

IPO - Center for User-System Interaction
Eindhoven, The Netherlands
c.h.b.a.v.dinther@tue.nl

Abstract

The perceptual relevance of changes to glottal-pulse parameters is studied. First, it is demonstrated that a distance measure based on excitation patterns can predict audibility discrimination thresholds for small changes to the R parameters of the Liljencrants-Fant (LF) model. Next, by using this measure the perceptual relevance of the LF parameters is quantified. Results are presented for a number of sets of glottal-pulse parameters that were taken from literature, representing distinct voice qualities.

1. Introduction

Glottal-pulse models have been of interest for quite some time. On the one hand, they provide insight in human speech production. On the other hand, they can help to increase the quality of synthesised speech [1]. Although much attention has been paid to relations between parameters of these models and voice quality, the perceptual relevance of these parameters has hardly been studied. In [2] the spectral relevance of glottal-pulse parameters was investigated. The consequences of small changes to the LF parameters for the resulting harmonic magnitude spectrum were considered. A natural extension is to apply the same type of analysis to a distance measure in a perceptual space. This should bring a better understanding of the perceptual difference between two sets of R parameters.

In order to analyse the perceptual relevance of the LF parameters, a perceptual distance measure is required. In [3], the L^2-norm between the excitation patterns, further denoted as the excitation pattern distance, was tested, together with other auditory measures such as the partial loudness measure, for arbitrary changes to the spectrum of stationary vowels. All these measures are based on an auditory model presented in [4]. It was shown that the partial loudness measure and the excitation pattern distance are equally appropriate for predicting audibility discrimination thresholds. Because of its mathematical properties, in this paper only the excitation pattern distance is used.

The excitation pattern distance is validated for application to speech stimuli. The measure is then used to analyse the perceptual relevance of the R parameters for the stationary Dutch vowels /a/, /i/ and /u/. In [2] it was found that the LF model, which, given the fundamental period, is a three-parameter model, actually operates as a one- or a two-parameter model. The results presented in this paper show a similar behaviour in the perceptual domain. It is shown that the glottal-parameter sets can be plotted as simple functions of one parameter, which indicates that all the parameter sets seem to be points on one trajectory in the glottal-pulse parameter space. When the perceptual relevance of the R parameter variations increases, only changes in one direction (or at most two) seem to remain perceptually relevant.

The outline of this paper is as follows. Section 2 briefly discusses the LF model. Section 3 presents the distance measure and Section 4 validates it for application to speech stimuli. The perceptual relevance of the parameters of the LF model is studied in Section 5 for a number of sets of estimated glottal-pulse parameters that were taken from literature. Section 6 presents the conclusions.

2. The Liljencrants-Fant model

The LF model has become a reference model for glottal-pulse analysis. Figure 1 shows an example of the glottal-pulse time derivative $g'(t)$ and introduces the specification parameters T_0, T_p, T_e and T_a. The length of a glottal cycle is $T_0 = 1/F_0$,

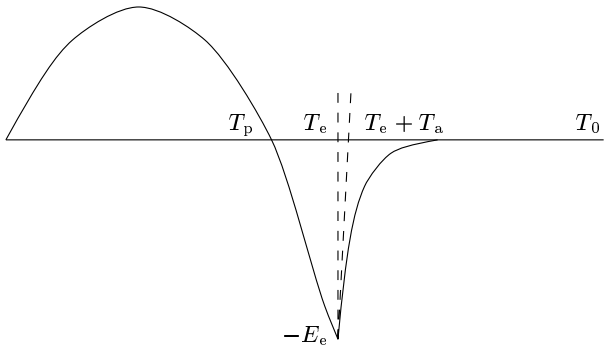

Figure 1: *Time derivative of the glottal-pulse.*

with F_0 being the fundamental frequency. The maximum airflow of the glottal-pulse occurs at T_p and the maximum excitation with amplitude E_e occurs at T_e, which corresponds to the instant when the vocal cords collide. The interval before T_e is called the *open phase*. The interval with approximate length $T_\mathrm{a} = E_\mathrm{e}/g''(T_\mathrm{e})$ just after the instant of maximum excitation is called the *return phase*. The interval between $T_\mathrm{e} + T_\mathrm{a}$ and the end of the glottal cycle is called the *closed phase*. During this phase the vocal folds have reached maximum closure and the airflow has reduced to its minimum. The minimum airflow is often referred to as leakage. Here we assume that there is no leakage, therefore $g(0) = g(T_0) = 0$. The airflow in the return phase is generally considered to be of perceptual importance, because it determines the spectral slope of the corresponding vowel. The parameters T_0, T_p, T_e and T_a are called the T-parameters. In this paper we will use the related set of R-parameters, which are defined as $R_\mathrm{o} = T_\mathrm{e}/T_0$,

[1] Also affiliated with Philips Research Laboratories, Eindhoven.

$R_{\mathrm{k}} = (T_{\mathrm{e}} - T_{\mathrm{p}})/T_{\mathrm{p}}$ and $R_{\mathrm{a}} = T_{\mathrm{a}}/T_0$. The parameters R_{o} and R_{a} denote the relative duration of the open phase and the return phase, respectively. The parameter R_{k} quantifies the symmetry of the glottal-pulse. The shape of the glottal pulse according to the LF model is fully specified by the parameters R_{a}, R_{k} and R_{o}. The R-parameters span a 3-dimensional subspace $\mathcal{R} \subset \mathbb{R}^3$ with elements $\mathbf{r} := [\mathrm{R_a}, \mathrm{R_k}, \mathrm{R_o}]^{\mathrm{T}} \in \mathcal{R}$.

3. The excitation pattern distance

The distance measure described in this section is based on a recently proposed 5-stage model for computing loudness [4]. For the implementation of the excitation patterns, only the first three stages are needed. The first two stages model transfer functions from the free field to the eardrum and through the middle ear. In the third stage, the excitation pattern of a given sound is calculated from the effective spectrum reaching the cochlea. According to [5], excitation patterns can be thought of as the distribution of "excitation" evoked by a particular sound in the inner ear along a frequency axis. In terms of a filter analogy, the excitation pattern represents the output level of successive auditory filters as a function of their centre frequencies. The auditory filter represents frequency selectivity at a particular centre frequency, where the filter shape varies with the input level. The excitation pattern is generally presented as a function of the ERB-rate rather than as a function of frequency. ERB stands for Equivalent Rectangular Bandwidth. The ERB-rate is a value on the so-called ERB scale. On this scale, the auditory filters are uniformly spaced. The excitation patterns are continuous functions on an interval in $\mathbb{R}$. In this paper the excitation patterns are calculated on an ERB-rate interval $(0, 40)$. Let

$$\xi : S \longrightarrow E \tag{1}$$

be the mapping of the N-dimensional spectral space $S \subset \mathbb{R}^N$ to the infinite excitation pattern space E. Here, the harmonic amplitude spectrum of a signal is represented as a point $(s_1, s_1, \ldots, s_N) \in S$, where N is the number of harmonics and $\{s_1, \ldots, s_N\}$ are the amplitudes of the harmonics of the spectrum of a signal. E is a subspace of $L^2(0, 40)$, where $L^2(0, 40)$ is the Banach space of Lebesgue integrable functions on the interval $(0, 40)$ under the norm $\| \quad \|$ given by

$$\|f\| := \Big(\int_0^{40} |f(x)|^2 dx \Big)^{\frac{1}{2}}. \tag{2}$$

In this paper, the integral is numerically approximated by a sum with $f(x)$ sampled at steps of 0.1 ERB. The excitation pattern distance is quantified by $\|e_1 - e_2\|$ for $e_1, e_2 \in E$.

The excitation pattern distance will be used to determine the perceptual relevance of small changes to the R parameters. Let

$$\phi_v : \mathcal{R} \longrightarrow S \tag{3}$$

be a mapping from the glottal-pulse parameter space to the spectral space S for a particular vowel v. In order to obtain the excitation pattern of a vowel v and a parameter set $\mathbf{r} \in \mathcal{R}$ the following mapping

$$\psi_v := \xi \circ \phi_v : \mathcal{R} \longrightarrow E \tag{4}$$

is used. The excitation pattern of a particular vowel v and parameter set $\mathbf{r}$ is then denoted by $\psi_v(\mathbf{r})$.

4. Validation of the excitation pattern distance

This section describes an experiment which has been done to validate whether the excitation pattern distance can be used to estimate audibility discrimination thresholds for small changes to the R parameters.

Stimuli. The stationary vowels /a/ and /i/ were synthesised with a source-filter model. The number of harmonics in this paper was N=36. The fundamental frequency of the vowels was F_0=110Hz. The sampling frequency was 8kHz. Two parameter sets $\mathbf{r}_1 := [0.03, 0.31, 0.56]^T$ and $\mathbf{r}_2 := [0.01, 0.33, 0.68]^T$ were varied in orthogonal directions in $\mathcal{R}$. In Table 1, a scheme is presented of the six resulting parameter conditions. The symbols +/− indicate if the specific R-parameter was increased/decreased.

Table 1: *Parameter conditions showing the directions of modifications to the R parameters.*

	Conditions and corresponding modifications		
	Direction of modification		
	/a/, $\mathbf{r}_1$	/a/, $\mathbf{r}_2$	/i/, $\mathbf{r}_1$
R_{o}	− (cond. 2)		+ (cond. 6)
R_{a}	+ (cond. 1)	− (cond. 3)	
R_{k}		+ (cond. 4)	− (cond. 5)

It is generally accepted that amplitude changes in the spectrum of harmonic sounds are more detectable than phase changes. For complex tones with a fundamental frequency beyond 150Hz, the maximal effect of phase on timbre is smaller than the effect of changing the slope of the amplitude pattern by 2 dB/oct [6]. It was tested whether phase changes could be ignored for complex tones with a fundamental frequency of 110Hz. For that purpose, two signal conditions were introduced. In one signal condition the phase was regular, i.e. the phase was allowed to change with the R parameters. In the other signal the phase was fixed, i.e. after modifying the signal the phases were adjusted to those of the reference signal. In total there were 2 (signal conditions) × 6 (parameter conditions) = 12 conditions. In the odd numbered conditions the signals have regular phase.

Experimental data. For each condition we took the norm of the excitation patterns of the modified sound with respect to the excitation patterns of the reference sound at which the subjects were able to discriminate between the reference sound and the modified sound.

Four subjects (EK, MB, MS and RD) participated in the experiments. All were young adults with normal hearing and no reported history of hearing impairment. The stimuli were presented diotically over headphones at a level of approximately 55 dB SPL. A 3IFC adaptive procedure as described in [3] was used to obtain thresholds. For each experimental condition the mean of four single-run estimates was taken as the final estimate of the distance for each subject. In Figure 2 the means are given for each subject, where the data of EK, MB, MS and RD are indicated with a square, circle, triangle and a star, respectively.

Results. The data presented in Figure 2 show that the thresholds of the subjects lie between 1.5 and 10 dB. The threshold values are a little higher than the results in [3]. The variation

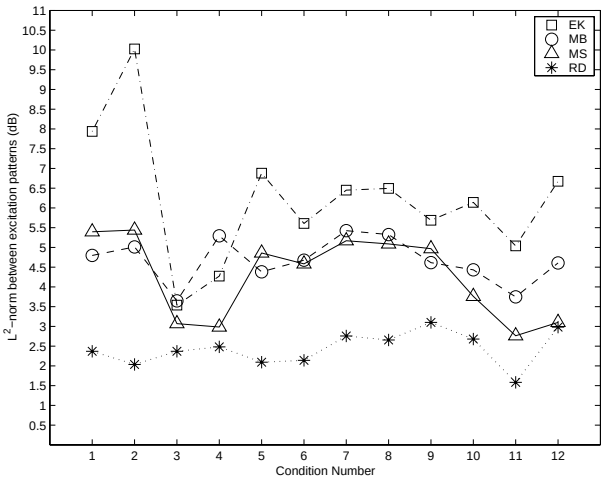

Figure 2: *Data of the obtained thresholds in terms of the excitation pattern distance.*

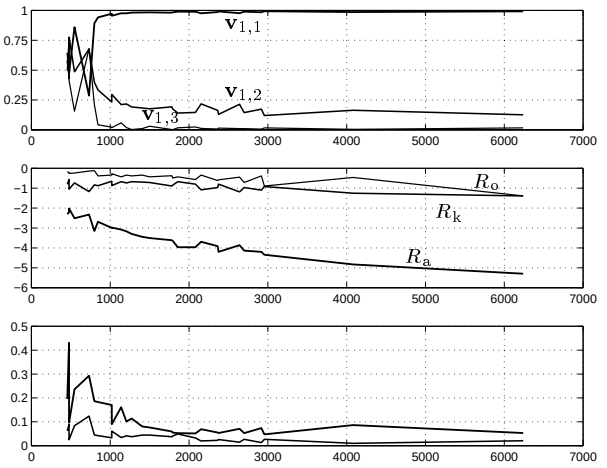

Figure 3: *Results for the vowel /a/. Top panel: entries of* $\mathbf{v}_1$ *as functions of* $\sqrt{\lambda_1}$*. Middle panel: The natural logarithm of the R parameters as functions of* $\sqrt{\lambda_1}$*. Bottom panel:* $\sqrt{\lambda_2/\lambda_1}$ *(upper line) and* $\sqrt{\lambda_3/\lambda_1}$ *(bottom line) as functions of* $\sqrt{\lambda_1}$*.*

in thresholds is partly caused by overall differences between the subjects with thresholds for the most sensitive subject being on average a factor 2 lower than those of the least sensitive subject. Across conditions, thresholds vary no more than about a factor 2 for each subject, indicating that the distance measure gives a reasonable indication of the subjects' sensitivity. Comparing the conditions with regular and fixed phases, it can be stated that there appear no systematic phase effects. It was concluded that the excitation pattern distance measure can be used for application to speech stimuli.

5. Perceptual relevance of the R parameters

The effects of small variations $\mathbf{h} \in [-10^{-4}, 10^{-4}]^3 \subset \mathbb{R}^3$ of a fixed set $\mathbf{r}$ were studied. The perceptual difference between the two sets **r+h** and **r** was quantified by the distance $\|\psi_v(\mathbf{r}+\mathbf{h}) - \psi_v(\mathbf{r})\|$ between the corresponding excitation patterns. The second-order approximation

$$\|\psi_v(\mathbf{r}+\mathbf{h}) - \psi_v(\mathbf{r})\|^2 \approx \frac{1}{2}\mathbf{h}^\mathrm{T} H(\mathbf{r})\mathbf{h}, \qquad (5)$$

is used with $H(\mathbf{r})$ the positive-definite 3×3 Hessian matrix of $\|\psi_v(\mathbf{r}+\mathbf{h}) - \psi_v(\mathbf{r})\|^2$. Let $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ denote the eigenvalues of $H(\mathbf{r})$ and $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ the corresponding orthonormal eigenvectors. The entry $\mathbf{v}_{i,1}$ corresponds to the parameter R_a, the entry $\mathbf{v}_{i,2}$ to the parameter R_k and the entry $\mathbf{v}_{i,3}$ to the parameter R_o. The eigenvalue λ_1 determines the maximum sensitivity of the distance measure as function of a variation of $\mathbf{r} \in \mathcal{R}$ in one direction and thus it quantifies the maximum perceptual relevance. The square root of λ_1 is used because (5) is a quadratic form. The ratios $\sqrt{\lambda_2/\lambda_1}$ and $\sqrt{\lambda_3/\lambda_1}$ determine the relative contribution to the perceptual relevance in directions orthogonal to $\mathbf{v}_1$. The numbers $\sqrt{\lambda_2/\lambda_1}$ and $\sqrt{\lambda_3/\lambda_1}$ quantify to what extend the LF model operates as a three-, two- or one-parameter model. The eigenvector $\mathbf{v}_1$ of $H(\mathbf{r})$ determines the direction of the maximum perceptual sensitivity according to the excitation pattern distance. Likewise, the eigenvector $\mathbf{v}_3$ of $H(\mathbf{r})$ determines the direction of the minimal perceptual sensitivity.

The eigenvalues and eigenvectors of $H(\mathbf{r})$ for the Dutch

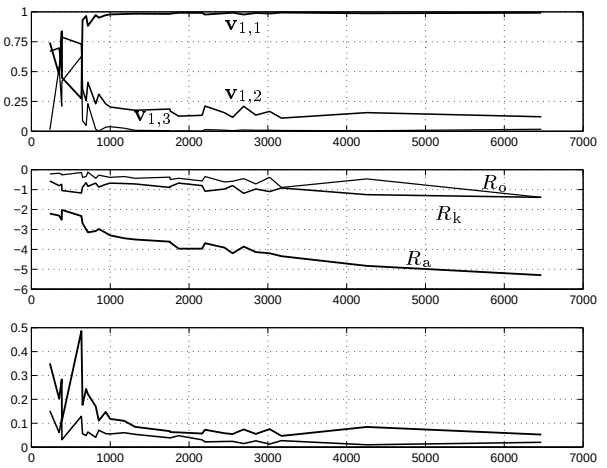

Figure 4: *Results for the vowel /i/. See Figure 3 for a description of the panels.*

vowels /a/, /i/ and /u/ were computed for 27 glottal-pulse parameter sets, which were taken from the references [7, 8]. The sets correspond to different voice qualities. The same glottal-pulse parameter sets were used as in Table 1 of [2]. The relative approximation error of (5) was calculated for entry 15 of Table 1 of [2]. On average it was 0.34% and maximally 5.96%.

Figures 3-5 show results obtained for the vowels /a/, /i/ and /u/. The top panels show the absolute values of the entries of the eigenvector $\mathbf{v}_1$ as functions of the eigenvalue $\sqrt{\lambda_1}$. The middle panels show the natural logarithms of the R parameters as functions of $\sqrt{\lambda_1}$. The bottom panels show the ratios $\sqrt{\lambda_2/\lambda_1}$ and $\sqrt{\lambda_3/\lambda_1}$ as functions of $\sqrt{\lambda_1}$.

There are strong similarities between the corresponding plots in Figures 3-5. This means that the differences in perceptual relevance between the R parameters of different vowels are small.

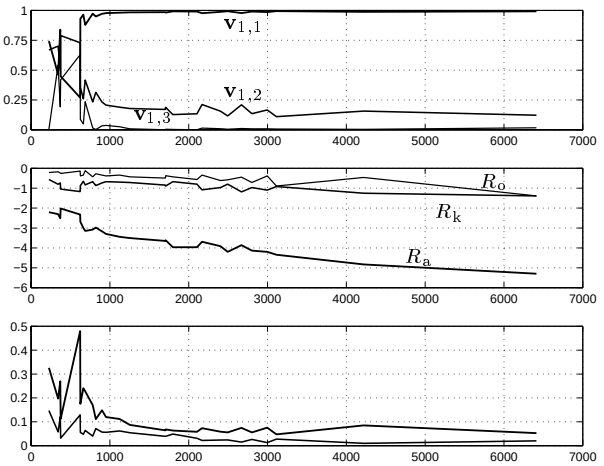

Figure 5: *Results for the vowel /u/. See Figure 3 for a description of the panels.*

It becomes clear from the figures that the R parameter sets can be plotted as simple functions of one parameter (i.e. $\sqrt{\lambda_1}$). This means that all the parameter sets seem to be points on one trajectory in the $\mathcal{R}$ space. Because the R parameter sets were taken from two different sources [7, 8] representing distinct voice qualities, it is surprising that these production parameters can be ordered along a trajectory which is a function of a perceptual parameter.

The top panels show that for, approximately, $\sqrt{\lambda_1} > 1000$, the $\mathbf{v}_1$ direction is nearly constant and corresponds to a variation mostly in the R_a and slightly in the R_k direction. The contribution of R_o is very low. For values $\sqrt{\lambda_1} < 1000$, there is no clear direction of maximal perceptual sensitivity and the behaviour is less consistent.

The effects of small R parameter variations in the directions $\mathbf{v}_2$ and $\mathbf{v}_3$ are compared with the effects of variations in the most significant direction $\mathbf{v}_1$. In the bottom panels of Figures 3-5, these effects are quantified by the square-root eigenvalues ratios $\sqrt{\lambda_2/\lambda_1}$ and $\sqrt{\lambda_3/\lambda_1}$. For, approximately, $\sqrt{\lambda_1} > 1000$, the ratios $\sqrt{\lambda_2/\lambda_1}$ and $\sqrt{\lambda_3/\lambda_1}$ are more or less constant. For $\sqrt{\lambda_1} < 1000$ the ratios and eigenvectors show more variation and jumps, which correspond to jumps in the R parameters. After approximating the R parameters in the middle panels with smooth curves and recalculating the eigenvalues and eigenvectors correspondingly, one can expect that the curves of the entries of the eigenvectors in the top panels will also be smooth.

The decision whether the LF model operates as a one-, two- or three-parameter model for these parameter sets, depends on a threshold for the square-root eigenvalues ratios. In [2] it was concluded that the LF model operates as a one or a two parameter model if an arbitrary choice of 10% for the threshold is made. Comparing this with the results in Figures 3-5, the LF model operates as a one parameter model for, approximately, $\sqrt{\lambda_1} \geq 1500$ and as a two- or three-parameter model for eigenvalues $\sqrt{\lambda_1} < 1500$. In any case, when the perceptual relevance of the R parameter variations increases, only changes in one direction (or at most two) seem to remain perceptually relevant.

6. Conclusions

In this paper it is demonstrated that a distance measure based on excitation patterns can predict audibility discrimination thresholds for small changes to the R parameters of the LF model. This measure was used to analyse the perceptual relevance of R parameters of the LF model.

The perceptual relevance of the R parameters can be quantified by one parameter. This is the square root of the maximum eigenvalue of the Hessian matrix in the second-order approximation of the excitation pattern distance.

It was found that the differences in perceptual relevance of the R parameters between the Dutch vowels /a/, /i/ and /u/ are small. The glottal parameters of these vowels can be plotted as simple functions of the parameter quantifying the perceptual relevance. It is surprising that these production parameters can be ordered along a trajectory which is a function of this perceptual parameter. It would be interesting to apply the same type of analysis to other, larger sets of glottal-pulse parameters and investigate if similar results can be obtained.

For the higher values of the parameter quantifying perceptual relevance the following observations were made. Firstly, the direction of maximum perceptual relevance becomes nearly constant and is mostly in the $R_\mathbf{a}$ and slightly in the $R_\mathbf{k}$ direction. Secondly, the LF model operates as a one- or two-parameter model, i.e. only changes in one direction (or at most two) seem to become perceptually relevant.

7. References

[1] Klatt, D.H. and Klatt, L.C., “Analysis synthesis, and perception of voice quality variations among female and male talkers”, J. Acoust. Soc. Amer., 87(2):820-856, 1990.

[2] Veldhuis, R.N.J., “The spectral relevance of glottal-pulse parameters”, Proceedings ICASSP, 873-876, Seattle, 1998.

[3] Rao, P., van Dinther, R., Veldhuis, R.N.J., Kohlrausch, A., “A measure for predicting audibility discrimination thresholds for spectral envelope distortions in vowel sounds”, accepted for J. Acoust. Soc. Amer., 2001.

[4] Moore, B.C.J., Glasberg, B.R., Baer, T., “A model for the prediction of thresholds, loudness and partial loudness”, J. Audio Engin. Soc., 45:224-240, 1997.

[5] Moore, B.C.J., Frequency Selectivity in Hearing, Academic Press Inc. London (LTD), 1986.

[6] Plomp, R., Steeneken, H.J.M., “Effect of phase on the timbre of complex tones”, J. Acoust. Soc. Amer., 46:409-421, 1968.

[7] Childers, D.G., Lee, C.K., “Voice quality factors: Analyses synthesis and perception”, J. Acoust. Soc. Amer., 90(5):2394-2410, 1991.

[8] Karlsson, I., Liljencrants, J., “Diverse voice qualities: Models and data”, TMH/QPSR, 2/96, KTH, 1996.