ANALYTICAL DESCRIPTION OF THE PRODUCTION OF FORMATS IN HUMAN SPEECH

Damyan Damyanov
Technical University of Sofia, Faculty of Automation, Department for Industrial Automation,
Bulgaria, Sofia, Darvenitsa 1756,Bul. Kliment Ohridksi 8, block 9,room 9420
damyan_damyanov@tu-sofia.bg

Keywords: model of speech production, man-machine systems.

Abstract: For the purposes. of speech synthesis, biometrics , medical and psychological diagnosis, the functioning of the glottis during phonation has been studied many times. At present, a relatively trivial solution proposed by Fant has established itself, regardless of the actual purpose of the system, using the "pusle source - filter" model. The model of Fant allows the linear prediction method to perform reconstruction of the current form of the vocal tract and the excitation of glottal volume velocity. But the practice shows that the fluctuations of the speech tract due to psycho-physiological effect on the functioning of the facial muscles in most cases are negligible. Thus, they are below the accuracy, which the linear model allows, using approximation with a cascade of coaxial cylindrical sections of equal length and constant cross-section. This requires more complex algorithms, and thus additional information is extracted from the pattern of air volume velocity after glottis. In this study, it is to be shown, that the model of Fant actually allows depiction of the psycho-physiological changes in the spectral features of the speech signal without the use of additional models. For this purpose it is sufficient to analyze the relationships of the main parameters of the excitation pulse of the source with the frequency response of the filter. In the current practice, these correlations are not considered and the source and the filter are examined separately.

1 INTRODUCTION

Phonation is a process which takes place in the middle part of the larynx and sets up vibrations to the exhaled airstream. As a consequence of this process the air volume velocity after the glottis immediately shows vibrational patterns (Figure 1), which determines the pitch frequency of the speech signal (Тилков, Д., Бояджиев., Т 1990). Changing the tension of the vocal folds and the pressure during the speech production process, one gets the desired pitch frequency, needed for phonation of vowels and voiced consonants. In the formation of non-voiced consonants the vocal folds do not come near each other and their constellation is like at physiological breathing (Pickett J.M, 1982). In this case, the incoming air pressure from the lungs is modulated by the formation of turbulence and closure of the vocal tract. The spectra of the resulting sound sources are modified by resonances, the frequency of which depends on the time-varying shape of the throat and mouth, the location of the tongue and many more.

Modern study of acoustic phenomena in the oral-pharyngeal tract began in 1941 with the work of Chiba and Kajiyama (Chiba, T., Kajiyama M., 1941). The fundament of the present theory and practice is shaped in 1960 by Fant (Fant, G., 1960), and others. The systems for processing, transmitting and storing of speech signals use various methods and techniques, but they are all based on a model of the functioning of articulatory tract and the air volume velocity after the glottis. The purpose of the modelling process of speech production is not only to study its properties and specifics, but also to solve problems for its effective coding and transmission ,

Damyanov D.
ANALYTICAL DESCRIPTION OF THE PRODUCTION OF FORMATS IN HUMAN SPEECH.
DOI: 10.5220/0004785600790084
In *Proceedings of the Second International Conference on Telecommunications and Remote Sensing (ICTRS 2013)*, pages 79-84
ISBN: 978-989-8565-57-0

the automatic synthesis of speech and speech recognition, for biometrical applications and others.

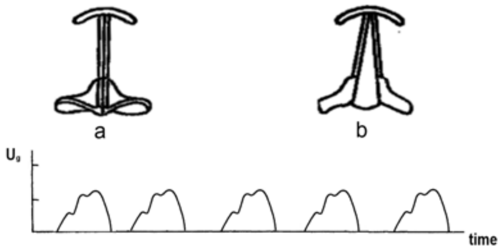

Figure 1: Air volume velocity immediately after the glottis for vowels and voiced consonants. Position of the vocal folds: a - phonation of vowels and voiced consonants, b - physiological breathing.

In every such model (Proakis, J.,2000), in one degree of another, the main components of articulatory tract in terms of the laws of acoustics are reflected (Figure 2)

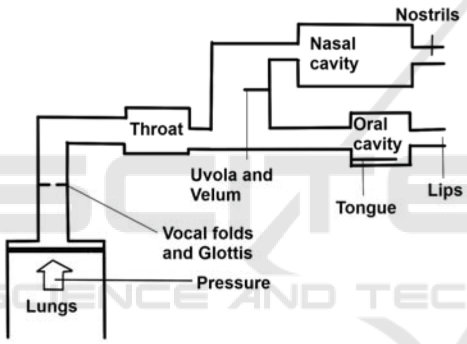

Figure 2: Main components of the acoustic articulatory tract.

These components include the excitation of the vocal cords, the time-varying shape of the vocal tract, radiation from the lips etc. Resonant phenomena in the speech production process are actually affected by losses in the walls of the tract, depending on the thermal conductivity, elasticity and friction. Significant influence is done by the nasal cavity with the absorption of certain frequency components of the spectrum (antiresonant phenomena). This complex nature of the process allows the development of many models, differing in structure and degree of relevance. Interestingly, despite the apparent complexity of the problem, it found a relatively trivial solution proposed by Fant (Fant, G., 1960), which has established itself in practice regardless of the actual purpose of the system.

The classical model of Fant separates the excitation from the shaping of individual sound components of speech, allowing approximately treatment of speech sounds as a linear system. Lungs, which provide air volume velocity, are presented as direct current power source. Their output is divided, corresponding respectively to the voiced and unvoiced parts of the signal. The pulse generator modulates the air volume velocity from periodical excitation of the vocal folds, and the noise generator models the formation of turbulence in the places, where the vocal tract changes its cross-section. The final speech signal is obtained after the components are added together and passed through a filter with linear transfer function, which models the articulatory tract.

In modern systems, this model has found its discrete implementation (Figure 3).

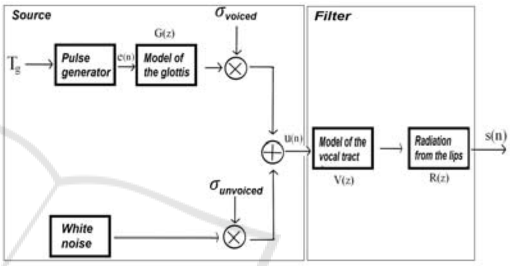

Figure 3: Digitalized model of Fant.

Generating a voiced speech segment after the z-transform of a sampled speech signal is described by the equation:

$$S(z) = U_g(z)V(z)L(z) \tag{1}$$

$$U_g(z) = \sum_{n=0}^{\infty}(z^{-K})^n G(z)\sigma_{son} = \frac{\sigma_{son}}{1 - z^{-K}}G(z) \tag{2}$$

It is assumed, that the sampling period is T=1, and the period of the pitch frequency is a multiple of it Tp = KT (Damyanov, D., Galabov, V., 2012a). The transfer function of the model of the glottis, the model of the vocal tract and the lips radiation can be combined in a common transfer function:

$$H(z) = G(z)V(z)L(z) \tag{3}$$

And the speech production process can be written as

$$S(z) = E(z)H(z) \tag{4}$$

$$S(z) = Z\{s(nT)\}, \quad s(nT) = s(t)\big|_{t=nT} \tag{5}$$

Where E(z) is the z-transformed pulse excitation of the glottis. Adoption of certain hypothesis about the spectral properties of the glottis, of the vocal tract and the lips radiation (Fant, G., 1960) allows the use of an all-pole filter:

$$H(z) = \frac{1}{1 + \sum_{i=1}^{M} a_i z^{-i}} \tag{6}$$

Finding the coefficients of this filter is a task, successfully solved in theory and practice. (Proakis, J.,2000, Trashlieva, V., Puleva, T., 2011)

2 MOTIVATION

Using a process of speech production of type source-filter, based on the model of Fant, many of the tasks, associated with the processing of speech signals are successfully solved. In this study, it is to be shown, that despite their simplicity, these models are able to explain some phenomena of the speech production, observed in the changes of current psycho-physiological state of humans. These include changed rations of the amplitudes of formant frequencies (less harsh or ringing voice), impaired understandability, etc. For their description and analysis of the functioning of the glottis during phonation has been studied many times, in order to create relevant model, especially for the purposes of speech synthesis, medical and psychological diagnosis, and biometrics. Much of the testing methods are borrowed from medicine in order to achieve greater accuracy, or to make measuring easier. All of the above methods do not work directly with the desired variable, but with some other, depending on the desired one. Given the complexity of making laryngoscopic and laryngoendoscopic studies, and especially the need for medical intervention, and because of the significant impact of these methods on the overall process of speech production, a great number of indirect methods have been developed. These include methods, base on electromyography - registering the signals of muscle activity in the throat while speaking, glottography - working with the impedance of the neck in the plane of the glottis, which largely depends on the cross-section of the air volume velocity after the glottis. Despite numerous studies, currently there is no unified theory that explains the process of these changes in the spectral characteristics of the speech signal. Most authors seek the cause of these phenomena in the psycho-physiological change in current form of the articulatory tract. The model of Fant allows the linear prediction method to perform reconstruction (with some degree of approximation) of the current form of the tract and the excitation of acoustic volume velocity after the glottis. It is expected, that the modelling of the speech tract provides conditions for obtaining relevant information about the functioning of the facial muscles and the corresponding parts of the nervous system, which are extremely sensitive to current psycho-physiological state of the individual (Ekman, P, Friesen, W., 1978). But the practice shows, that the fluctuations of the speech tract due to psycho-physiological effect on the functioning of the facial muscles in most cases are too small . Thus they are below the accuracy, which the linear model allows, using approximation with a cascade of coaxial cylindrical sections of equal length and constant cross-section (Pfister, B., Kaufmann, T., 2008).

This requires more complex algorithms and thus additional information is extracted from the pattern of air volume velocity after glottis. The air volume velocity gives information about the current psycho-physiological condition in two ways - first, through the entire process of speech production, concerning the higher nervous activity, and the second by the functional state of the autonomic nervous system, including the overall muscular tonus (Reuter-Lorenz, Patricia A., et.al., 2010, Kalat, James W. 2012). For the extraction of this information, the use of relatively complex models is required, describing the tension on the vocal folds, the mechanism of stretching and vibrating, the pressure of the incoming air flow from the lungs, etc. In this study, it is to be shown, that the model of Fant actually allows depiction of the psycho-physiological changes in the spectral features of the speech signal without the use of additional models. For this purpose, it is sufficient to analyze the relationships of the main parameters of the excitation pulse of the source with the frequency response of the filter. In the current practice, these correlations are not considered and the source and the filter are examined separately. For readability purposes, the following simplifications will be made that do not change the generality of the study:

- The influence of the transfer function of the glottis and the lips radiation will be neglected, since they don't affect the formant frequencies, but only their decay ;

- The excitation of the source will be considered as a sequence of rectangular pulses;
- The attenuation of formant frequencies will be ignored, i.e. poles of the filter of the vocal tract will lie of the unit circle;

3 PRODUCTION OF FORMANTS IN HUMAN SPEECH

In this study, it is assumed, that the model of the vocal tract is of second order:

$$H(s) = \frac{k_1\,\omega_1^2}{s^2 + \omega_1^2} \tag{7}$$

i.e. the signal will contain only one formant with circular frequency ω_1 . The excitation pulses are of type (Damyanov, D., Galabov, V., 2012b):

$$e(t) = \begin{cases} 1, & (m-1)T_g \le t < t_{open\_glottis} + (m-1)T_g \\ 0, & t_{open\_glottis} + (m-1)T_g \le t < mT_g \end{cases} \tag{8}$$

$$m = \overline{1, N_{imp} - 1}$$

where T_g is the period of the pitch frequency , and $t_{open\_glottis}$ is the duration of the phase of open glottis. The output signal for the period of the pitch frequency (m=1) is:

$$s^{1F}_{open\_glottis}(t) = AO^{1F}_{open\_glottis} - A^{1F}_{open\_glottis}\sin(\omega_1 t + \varphi^{1F}_{open\_glottis}) \tag{9}$$

for $t_{open\_glottis}$, i.e. in the phase of open glottis, and:

$$s^{1F}_{closed\_glottis}(t) = A^{1F}_{closed\_glottis}\sin(\omega_1 t + \varphi^{1F}_{closed\_glottis}) \tag{10}$$

for $t \ge t_{open\_glottis}$, i.e. in the phase of closed glottis, where:

$AO^{1F}_{open\_glottis} = k_1$ is the DC component of the signal in the phase of open glottis

$A^{1F}_{open\_glottis} = k_1$ and

$A^{1F}_{closed\_glottis} = 2k_1 \sin(\frac{\omega_1 t_{open\_glottis}}{2})$ are the amplitudes of the signal in the phases of open and closed glottis

$\varphi^{1F}_{open\_glottis} = \frac{\pi}{2}$ and

$\varphi^{1F}_{closed\_glottis} = 2\pi - \frac{\omega_1 t_{open\_glottis}}{2}$ are the angular phases of the signal in the phases of open and closed glottis

$\omega_1 = 2\pi f_1$ are the circular frequency, which corresponds to the formant frequency f_1 .

We can observe the following important features:

- In the two phases - of open and closed glottis - the formant frequency is determined only on the coefficient of plain gain of the filter
- The amplitude of the signal in the phase of closed glottis depends again on this coefficient, but also in a complicated way on the ratio of duration of the preceding phase of open glottis to the period of the formant frequency.
- The phase shift of the signal depends in a similar way on the above stated variables.

Practically, this means that without changes in the filter parameters, i.e. the geometry of the vocal tract, changes in length of the phase of open glottis can increase or decrease the formant frequencies of the speech signal (Damyanov, D., Galabov, V., 2012a). For better understanding of the paradigm, a dimensionless coefficient is introduced, which is proportional to the duration of the phase of open glottis to the period of the formant frequency:

$$r_{\omega_1 t_{open\_glottis}} = \omega_1 t_{open\_glottis} \tag{11}$$

In terms of amplitude of the signal before and after closure of the glottis the coefficient can be defined:

$$r_{A^{1F}_{open\_glottis} A^{1F}_{closed\_glottis}} = \frac{A^{1F}_{open\_glottis}}{A^{1F}_{closed\_glottis}} \tag{12}$$

where the $t_{closed\_glottis} = T_g - t_{open\_glottis}$ is the duration of the phase of closed glottis during the period of the pitch frequency (Damyanov, D., Galabov, V., 2012b). The functional relationship between these two coefficients is:

$$r_{A^{1F}_{open\_glottis} A^{1F}_{closed\_glottis}} = f\left(r_{\omega_1 t_{open\_glottis}}\right) = \left|2\sin\left(\frac{\omega_1 t_{open\_glottis}}{2}\right)\right| \qquad (13)$$

It can be seen, that varying the duration of the phase of open glottis, without changing the filter parameters, i.e. the geometry of the vocal tract, for a given amplitude of the generated signal in the phase of closed glottis can take any value from zero to two times the amplitude in the phase of open glottis (Damyanov, D., Galabov, V., 2012c).

This effect becomes more interesting, if the speech segment contains several periods of the pitch frequency. Then, besides the coefficient $r_{\omega_1 t_{open\_glottis}} = \omega_1 t_{open\_glottis}$, the ratio of the amplitudes will be affected by the ratio of the length phase of closed glottis during the period of pitch frequency:

$$k_{full\_imp} = \frac{t_{open\_glottis}}{T_g} \qquad (14)$$

In this case one can obtain relatively complex analytical dependencies. For example the relationship of the amplitudes of the signal in the phase of closed glottis to the phase of open glottis in the second period of the pitch frequency is given by

$$r_{A^{1F}_{open\_glottis} A^{1F}_{closed\_glottis}} = f(r_{\omega_1 t_{open\_glottis}}) = \qquad (15)$$
$$\left[4k_1 \sin\left(\frac{r_{\omega_1 t_{open\_glottis}}}{2}\right)\cos\left(\frac{r_{\omega_1 t_{open\_glottis}}}{2k_{full\_imp}}\right)\right]$$
$$\left[\left(k_1 \sin(r_{\omega_1 t_{open\_glottis}}) + k_1 \sin\left(2\pi - \frac{r_{\omega_1 t_{open\_glottis}}}{k_{full\_imp}}\right)\right)^2 + \left(-2k_1 \sin^2\left(\frac{r_{\omega_1 t_{open\_glottis}}}{2}\right) + k_1 \sin\left(\frac{r_{\omega_1 t_{open\_glottis}}}{k_{full\_imp}} - \frac{\pi}{2}\right)\right)^2\right]^{-\frac{1}{2}}$$

If we assume that:

$$\begin{aligned} &numbits = total\ number\ of\ bits \\ &\quad from\ the\ binary\ presentati on\ of\ N \end{aligned} \qquad (16)$$

Then, for the N-th period of the pitch frequency the dependencies are as follows:

$$A^{1F}_{closed\_glottis} = 2k_1' \sin\left(\frac{\omega_1 t_{open\_glottis}}{2}\right) \qquad (17)$$
$$\left\{\begin{aligned} &\sin\Bigg(\omega_1 t - \frac{\omega_1 t_{open\_glottis}}{2}\Bigg(1 + \frac{1}{k_{full\_imp}} \\ &\sum_{sab=2}^{numbits-1} 2^{numbits} + 2^2 BSB\Bigg)\Bigg) BFB \\ &+ \sum_{sab=2}^{numbits}\left(\prod_{ampl=2}^{sab} 2\cos\left(\frac{\omega_1 t_{open\_glottis}}{k_{full\_imp}} 2^{(ampl-3)}\right)\right) \\ &+ \sin\Bigg(\omega_1 t - \frac{\omega_1 t_{open\_glottis}}{2}\Bigg(1 + \Bigg(2^{sab-1} - 2^0 + 2^{sab+1} \\ &\Bigg(\sum_{\substack{sub=sab+1 \\ sab+1<numbits}}^{numbits} BBS\Bigg)\Bigg)\Bigg)\Bigg) BSB \end{aligned}\right\}$$

where

$$BSB = \begin{cases} 0, \text{ if the } sab-th \text{ significan t bit} \\ \quad \text{from the binary presentati on of } N \\ \quad \text{is equal } 0 \\ 1, \text{ else} \end{cases} \qquad (18)$$

$$BFB = \begin{cases} 0, \text{ if the first bit} \\ \quad \text{from the binary presentati on of } N \\ \quad \text{is equal } 0 \\ 1, \text{ else} \end{cases} \qquad (19)$$

$$BSS = \begin{cases} 0, \text{ f the } sub-\text{significan t bit} \\ \quad \text{from the binary presentati on of } N \\ \quad \text{is equal } 0 \\ 1, \text{ else} \end{cases} \qquad (20)$$

$$BBS = \begin{cases} BSS, \text{ if } sab = numbits \\ BSB + numbits - sab - 1, \text{ else} \end{cases} \qquad (21)$$

and

$$A^{1F}_{open\_glottis} = k_1' + A^{1F}_{closed\_glottis}((N-1)T_g) \qquad (22)$$
$$-\sin\left(\omega_1 t - \frac{\omega_1 t_{open\_glottis} N}{k_{full\_imp}} + \frac{\pi}{2}\right)$$

One can make the following conclusion:

By changing the ratio of the duration of the phase of close glottis to the length of the pitch frequency and without changing the geometry of the vocal tract, one can generate speech segments, in which the formant amplitude for each subsequent period increases, decreases, or does not change much. This can be seen in the two parts of figure 4. In the figure, a speech signal, generated with the model of Fant is resented. It contains one formant, which has frequency 420 Hz. The pitch is 199 Hz, and the plot contains 9 pitch periods. The signal in the first part has $t_{open\_glottis} = 39\ ms$, and the second part has $t_{open\_glottis} = 25\ ms$.

4 CONCLUSIONS

The study shows that the model of Fant adequately describes many phenomena of the speech production process, that are known form theory and practice. The reason for this probably lies in the genesis of the model. A "pulse source-filter" model is created for simplification of the process of analysis and parameterization of a specific implementation. This is accomplished, with the assumption, that the filter is practically independent of the source of excitation. This allows the implementation of effective methods and techniques for the evaluation of the filter. This approach gives excellent results in most cases - mainly in speech analysis, synthesis, coding and transmission. When the speech signals are used for medical, psycho-physiological, biometrical purposes however, it is customary to consider that the model of Fant is not sufficiently relevant. This requires the use of much more complex models and additional information sources. Recently, the problem becomes even more acute with the ever increasing demands on quality of processing and transmission of speech signals and their use in mobile devices. This paper shows, that the model can be made much more efficient without further complication. This is achieved, using the cumulative effect of some of the processes, that can be found in source and in the filter.

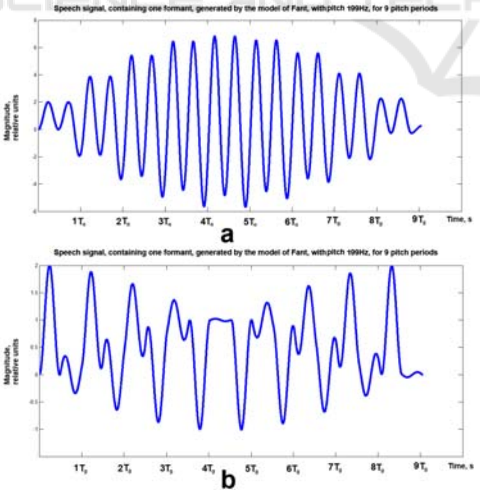

Figure 4: Speech signal, generated with the model of Fant, containing one formant, with pitch 199 Hz, 9 pitch periods, with formant frequency 420 Hz, and duration of the phase of open glottis 39 ms for a) and 25 ms for b).

REFERENCES

Chiba, T., Kajiyama, M.1995,*The vowel, Its Nature and Structure*. Tokyo-Kaiseikan, Tokyo 1995

Damyanov, D., Galabov, V., 2012a, *Characteristics of the model of Fant of second order on speech production*, Proceedings of the Technical University - Sofia, Volume 62, Issue 2, pp. 181-188, ISSN 1311-0829 , Sofia, 2012,

Damyanov, D., Galabov, V., 2012b, *On the Impact of duration of the phase of open glottis on the spectral characteristics of the phonation process*, Proceedings of the Technical University - Sofia, Volume 62, Issue 2, pp. 173-180, ISSN 1311-0829 , Sofia, 2012,

Damyanov, D., Galabov, V., 2012c, *Some effects of the assumption of an all-pole filter, used to describe processes of type "pulse source*, 1-st International Conference on Telecommunications and Remote Sensing, August, 29-30, pp. 139-145, ISBN 978-989-8565-28-0 , Sofia, 2012,

Ekman, P., Friesen W., 1978, *The Facial Action Coding System*, Consulting Psychologist Press, San Francisco. CA, 1978

Fant, G., 1990, *Acoustic Theory of Speech Production*, Mouton&Co, Hauge

Flannagan, J., 1992, *Speech analysis, Synthesis and Perception*, Springer, Berlin, 1992

Hayes, M., 1999, Schaum's Outline of Theory and Problems of Digital Signal Processing, Singapore, McGraw-Hill, 1999

Kalat, James W. 2012, *Biological Psychology*, Wadswoth. Cengage Learning, 10-th edition, Belmont, 2009

Pfister, B., Kaufmann, T., 2008, *Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung*, Springer Verlag, Heidelberg, 2008

Pickett, J.M. 1982, *The sounds of speech communication* , Univercity Park Press, Baltimore, 1982

Proakis, J.,2000, *Discrete Time Processing of Speech Signals*, New Jersey, JohnWiley&Sons, IEEE Press, 2000

Rabiner, L., Schafer R. 1992, *Digital processing of speech signals*, Prentice-Hall Inc, Engelwood Cliffs, New Jersey, 1992

Reuter-Lorenz, Patricia A., et.al., 2010, *The Cognitive neuroscience of mind: A tribute to Michael S. Gazzaniga*, MIT Press, April, London, 2010

Trashlieva, V., Puleva, T., 2011, *Model building for optimal administrative process management*, International Conference Automatics and Informatics'11, Bulgaria, 3-7.10.2011, pp B-263-B-266, ISSN 1313-1850, Sofia, 2011

Тилков, Д., Бояджиев., Т 1990, *Българска Фонетика*, Наука и изкуство София 1990