INVESTIGATING SOURCE-FILTER INTERACTION TO SPECIFY CLASSIC SPEECH PRODUCTION THEORY

Vera Evdokimova, Karina Evgrafova, Pavel Skrelin

Saint-Petersburg State University, Saint-Petersburg, Russia
postmaster@phonetics.pu.ru, evgrafova@phonetics.pu.ru, skrelin@phonetics.pu.ru

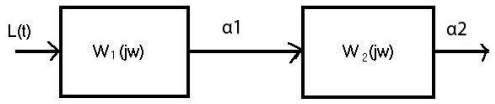
ABSTRACT

The paper is concerned with the specification and improvement of the traditional source-filter model of the human vocal tract proposed by G.Fant and analyzed by many scientists. The new method of recording the glottal wave synchronously with an output speech signal was employed to obtain the experimental material. The comparison of the recorded signals allowed analyzing the structure of the speech signal at different stages of its generation. As a result, the classic vocal tract model was specified by distinguishing a feedback component which formalizes the processes in the vocal tract as a complex acoustic nonlinear system. One of the functions of the component is to transform the acoustic energy from the articulation system upstream.
In the paper the recording method is described, perceptual experiment and the acoustic analysis results are presented.
Keywords: phonetics, phonation, voice source, feedback component, source-filter interaction, formants.

1. INTRODUCTION

The traditional approach to phonetic research of the vocal tract assumes that there are several successive stages of speech production which are initialization, phonation, articulation and radiance of speech signal. The initialization is an impact of muscular and pulmonary systems (lungs) and the power basis of the speech production process. The phonation stage provides the input signal to the filter component of the vocal tract. This input signal contains the fundamental frequency and its high harmonics. The voice signal goes through the filter component - a set of pharynx, nasal and oral cavities. The voice source signal is the strongest acoustic signal in the human vocal tract. Almost all the internal organs are parts of the bio-mechanical oscillating system that generates the voice signal. This signal is individual and optimized by nature [1],[2],[3],[4],[8]. The periodic sequence of lung pressure differences in larynx is called the glottal wave [5],[6]. The frequency of these pulses corresponds to the fundamental frequency in speech signal.

Figure 1: Dynamic system of the vocal tract consisting of two parts.



L(t) – air flow pressure from the respiratory apparatus (the lungs),
$W_1(jw)$ - frequency characteristics of the source component that includes trachea, larynx and the vocal chords,
$\alpha_1(t)$ - output acoustic signal of the source component that includes the pitch ant its high harmonics, also it includes a lot of other frequencies which were reduced on that stage,
$W_2(jw)$ - frequency characteristics of the articulation,
$\alpha_2(t)$ – speech signal.

The fact of the interaction between the two parts of the vocal tract does not make the classic linear source-filter theory completely consistent. Obtaining the voice source signal detached from the influence of the articulation system and analyzing its nature is an important up-to-date problem for different fields of speech science and speech technology.

There exist different voice source models that are applied to the majority of linguistic research and speech technology applications. The LF-model (Lilencrants and Fant) of the voice source was developed in the 80-s by G. Fant [5], [6]. It described the glottal wave as a sequence of pulses of the given shape. The frequency of these pulses is the fundamental frequency. Their shape is similar to the experimentally measured shape of glottal pulses. Comparing the model with the pattern showed that the voice signal can be modeled successfully by the derivative of glottal wave function.

The improved quality of the speech synthesis system based on the LF-model was caused by the fact that not only the pitch but also its high harmonics were taken into account. The basis of interference of voice and filter components was maintained in the model. The LF-model imitated the

voice signal and worked well for a predefined voice of text-to-speech synthesis system. However, it is more complicated to use it for real time analysis of voice. The determining of the LF-model parameters is a very complicated task which requires many calculations.

Apart from LF-model, there exist biomechanical models of the voice source and the vocal folds. Single-mass models could be more precisely termed single degree-of-freedom vibration models for the vocal folds because each vocal fold is modeled with a single mass-spring system. Generally, it is assumed that the mass-spring systems modeling each vocal fold are identical and the glottal flow is assumed to be the same on the either side of the glottal center-line [15], [12]. These models are of particular interest theoretically because they must explain the net work done on the vocal folds by air flow in 1 cycle in terms of asymmetries between opening and closing phases in air flow conditions. This is in contrast to the two-mass model which explains the energy input into the vocal folds using asymmetries in vocal fold geometry between opening and closing. The two-mass vocal fold model introduced by Stevens [4] consists of two pairs of masses, larger ones representing the inferior part of the vocal folds, and smaller ones representing the superior part of the vocal folds. The mechanism depends on the fact that the inferior and superior part of the vocal folds do not move together as a rigid body. The whole process of the excitation function computation is described in Chapter 2 of [4].

In the traditional models the filter cannot influence the source to produce new frequencies or change the overall energy level of the source. Titze [13], [14] showed that this assumption is generally not valid. However, under certain conditions it is an appropriate simplification. The traditional linear source–filter theory is encumbered with possible inconsistencies in the glottal flow spectrum, which is shown to be influenced by interaction. The source-filter interactions that involve changes in vocal fold vibration have been demonstrated by investigators [14], [16], [17], [7]. However, the data presented are sometimes fragmentary and inconsistent. The main Titze's goal was to determine the proportion of irregularities that are due to nonlinear source-tract interactions and to provide a theoretical framework for the bifurcation phenomena in vocal fold vibration with a nonlinear source-filter construct [13], [14]. As long as the dominant source frequencies lie well below the formant frequencies of the vocal tract, the source is influenced only in simple ways by the filter, mainly in terms of glottal flow pulse skewing and pulse ripple. This mild interaction occurs for most male adult speech, but greater interaction occurs for female and child speech, and even more for singing, where the fundamental frequency range spans more than two octaves and the lower partials of the source cross the formants. In these more intense interactions, bifurcations in the dynamics of vocal fold vibration can occur that may generate sudden F_0 jumps, subharmonic frequencies, or changes in the overall energy level at the source [13], [14].

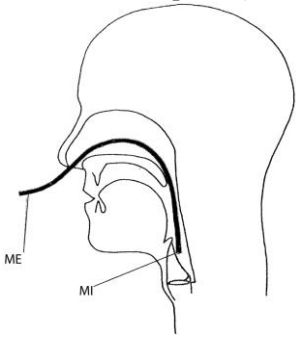
Our research was aimed at registering and analyzing the signal obtained through the microphone placed in the proximity of the vocal folds (Microphone Internal - MI) and comparing it with the output speech signal (Microphone External - ME). Besides, the non-linearity of the vocal tract system is considered. We hypothesized the obligatory presence of the feedback component. The goal of the experiment was to show that the energy from the articulation system reflects upstream. The main interest of the paper is the acoustic characteristics of the signals, mainly, the vowel formant structure.

2. METHOD

2.1. Equipment

The recordings were made in the recording studio. Multichannel recording system Motu Traveler and WaveLab program were used. The recordings had a sample rate of 44100 Hz and a bitrate of 16 bits. Two microphones were used during the recording. The capacitor microphone AKG HSC200 was placed in the output of the speakers mouth (ME). The miniature microphone QueAudio (d=2.3 mm, waterproof) was located in the proximity of the speaker`s vocal folds in the larynx (MI) with the use of special medical equipment. This procedure was performed by a phoniatrician.

Figure 2: Location of two microphones used for the recording (MI- internal microphone, ME – external microphone).



2.2. Subjects

The subjects of the experiment were male and female speakers. The female speaker pronounced four utterances of each of the 6 Russian vowels: /a/, /e/, /i/, /ɨ/, /o/ and /u/. For the male speaker, only two utterances per vowel were available.

3. PERCEPTUAL ANALYSIS

Two groups of informants (25 individuals) were involved into perceptual tests. The groups consisted of 5 participants who were experts in phonetics and 20 lay participants. The MI vowel stimuli were presented to informants in order to find out the way if a voice source system could be identified as a speech sound. The samples were organized on a random basis. The informants were asked to make judgments with respect to each stimulus and decide whether it could be identified as any of vowel phonemes.

Overall, the perceptual test did not confirm the presupposition that all the MI vowel stimuli should sound similar. It was based on the classic theory of speech production by G. Fant according to which these are formants that define the phonetic quality of a vowel.

However, the group of expert participants 1) distinguished and all the stimuli 2) recognized correctly practically all of them. The Russian vowels /i/ and /y/ which are quite similar were confused by some participants. The group of lay participants also answered that they could hear different vowels. However, not all of the stimuli sounded intelligible for them.

The MI vowels /a/, /e/, /i/ stayed most intelligible and were identified correctly in most instances. However, there were strong confusions of /i/ and /u/, /i/ and /y/ and /u/ and /y/. Besides, /a/ and /i/ vowels were often perceived as labialized.

4. ACOUSTIC ANALYSIS

The spectral characteristics of the vowels were obtained in order to compare the frequency constituents of both signals.

Figure 2: Spectral densities of the MI (dashed line) and ME (solid line) signals for the vowels /a/, /e/. Female speaker. **/a/** **/e/**

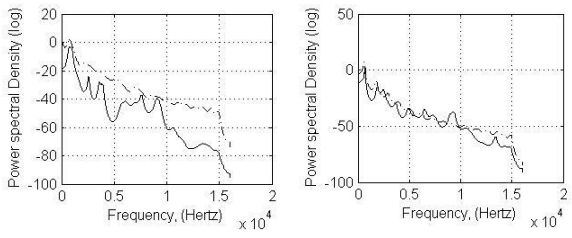


Fig 2 shows the difference in the frequency constituents for the two signals for vowels /a/ and /e/. The difference lies not only in the formant region but also in the higher frequency constituents. However, in this part of the research we were primarily interested in the formant structure of the vowels.

Fig. 3, 4 show the examples of spectral densities of the 6 Russian vowels. There are two curves in each picture. The solid line shows the spectral density of ME signal and the dashed line shows the spectral density of MI signal. The analysis of the vowel spectra shows that the signal from MI contains the frequency constituents of the vowel formants (resonance frequencies of the set of pharynx, nasal and oral cavities) However, the frequency constituents are weakened. It can be assumed that it is caused by the reflection of the acoustic energy from the articulation system backwards. As well as this the plots show that the signals can be very different for the two microphones. For example, see the plots for the vowel /ɨ/ (fig.3).

Figure 3: Spectral densities for vowels /a/, /e/, /i/, /ɨ/, /o/, /u/. Male speaker. The solid line shows the spectral density of ME signal and the dashed line shows the spectral density of MI signal.

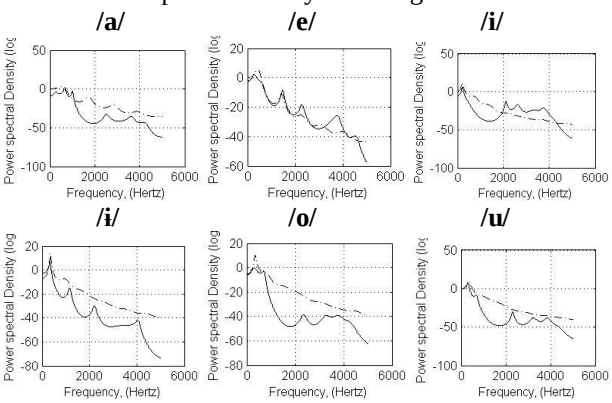
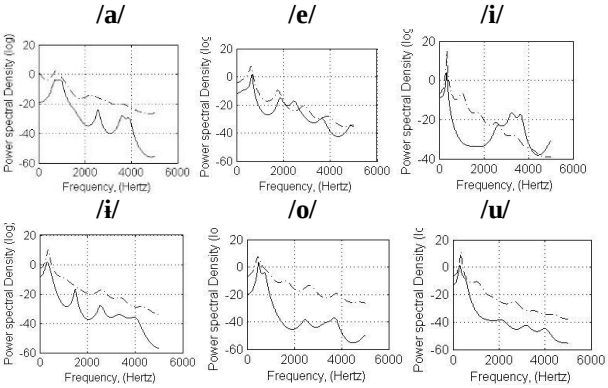


Figure 4: Spectral densities for vowels /a/, /e/, /i/, /ɨ/, /o/, /u/. Female speaker. The solid line shows the spectral density of ME signal and the dashed line shows the spectral density of MI signal.

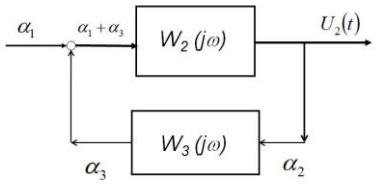


5. MODELLING

The human vocal tract is usually regarded as a unified dynamic system consisting of two concatenated parts which are the voice source and filter component which have their own dynamic characteristics. The both parts are non-separable and interact.

The analysis of the signal made by MI made it possible to correct this model by adding a feedback component.

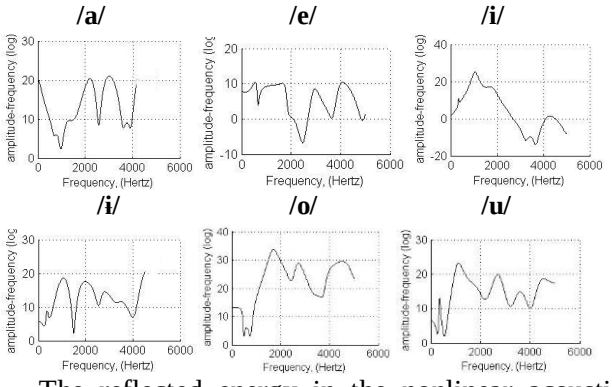
Figure 5: System of articulation of the human vocal tract with a feedback component.



α_1 – glottal wave,
α_2 – speech signal,
$U_2(t)$ – output speech signal,
α_3 – reflected backwards acoustic energy,
$W_2(jw)$ - frequency characteristics of the articulation,
$W_3(jw)$ - frequency characteristics of the feedback component.

Equivalent logarithmic amplitude-frequency characteristics of the feedback component (ELAFFS) for the above vowels are on the following plots:

Figure 6: Equivalent logarithmic amplitude-frequency characteristics of the feedback component (ELAFFC) for the above vowels /a/, /e/, /i/, /ɨ/, /o/, /u/. Female speaker.

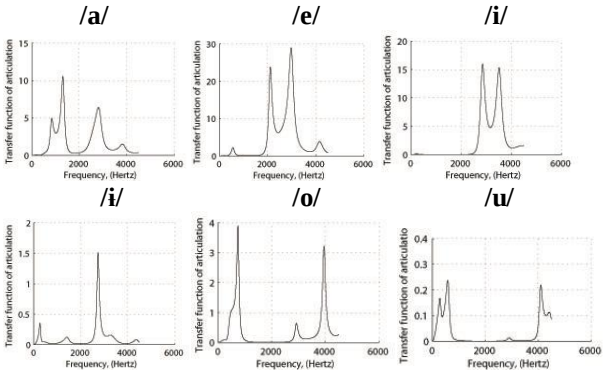


The reflected energy in the nonlinear acoustic system of the vocal tract has an influence on the work of the voice source and the glottal wave characteristics. Also this energy is being reflected again to the articulation system and its frequency constituents are changed.

Besides, there is a reverberation due to the flexible walls of the pharynx. The acoustic waves are reflected repeatedly. Yet the process can be regarded as a stationary one and is defined by a feedback component.

The coprocessing of several acoustic realizations with different levels of influence of the vocal tract parts helps to elaborate the methods of discrimination and modeling the transfer functions of filter component of the vocal tract for vowels and to obtain the formant structure. Formant positions were estimated for voice and supraglottal waves using the algorithm described in the paper [10].

Figure 7: Filter component transfer function of the stressed vowels /a/, /e/, /i/, /ɨ/, /o/, /u/ (30 ms). Female speaker. The formant structure is well-defined.



6. CONCLUSIONS AND DISCUSSION

The new method of synchronous recording of the speech signal by placing the microphone near the vocal folds allowed us to specify and improve the source-filter model. The results of the experiment confirmed the fact that the acoustic energy from the filter component reflects backwards. This justified the introduction of the feedback component into the model. The proposed approach allows automatic discriminating of the vowel formant structure by processing the real speech data. The constructed model of the filter part of the vocal tract completely corresponds to the basic phonetic laws. It adds the accuracy to the existing models of the speech production and can be used for solving the specific problems of speech technologies.

7. ACKNOLEDGEMENTS

This work has been carried out in the framework of SPbSU project n. 31.37.353.2015 (“The Phonetic Aspects of Speech Signal Synthesis with a High Degree of Naturalness”).

8. REFERENCES

[1] Bondarko L.V. Phonetics of Russian modern language, SPbSU, 1998 (in Russian).
[2] Kodzasov S.V., Krivnova O.F. Moscow. 2001. General Phonetics.

[3] Fant G. 1960. Acoustic Theory of Speech Production. Netherlands: Mouton.
[4] Stevens, K. 1998. Acoustic Phonetics. Cambridge, MA 02141: The MIT Press.
[5] Fant, G. 1997. The voice source in connected speech. Speech Communication, v. 22.
[6] Fant G., Liljencrants J., Lin Q. 1985. A four-parameter model of glottal flow. STL-QPSR, . 2-3.
[7] Flanagan J. L. 1968. "Source-system interaction in the vocal tract," Ann. N.Y. Acad. Sci.155, 9–17.
[8] Flanagan J. L. 1972. Speech Analysis, Synthesis, and Perception. Springer, New York.
[9] Bessekersky V.A., Popov E.P. 1972 Automatic control theory systems. Moscow, Nauka (in Russian).
[10] Evdokimova V.V. 2006. The use of vocal tract model for constructing the vocal structure of the vowels // SPECOM'2006, Saint-Petersburg, 25-29 June 2006, p. 210-214.
[11] Sergienko A.B. 2003. Digital signal processing. Moscow, 2003 (in Russian).
[12] Howe M. S., and McGowan R. S. 2010. "On the single-mass model of the vocal folds,"Fluid Dyn. Res. 42, 015001. doi: 10.1088/0169-5983/42/1/015001.
[13] Titze I. R. 2008. "Non-linear source-filter coupling in phonation: Theory," J. Acoust. Soc. Am. 123, 2733–2749. doi: 10.1121/1.2832337.
[14] Titze I. R., Riede T., and Popolo P. 2008. "Nonlinear source-filter coupling in phonation: Vocal exercises," J. Acoust. Soc. Am. 123, 1902–1915. doi: 10.1121/1.2832339.
[15] Zanartu M., Mongeau L., and Wodicka G. R. 2007. "Influence of acoustic loading on an effective single mass model of the vocal folds," J. Acoust. Soc. Am. 121, 1119–1129. doi: 10.1121/1.2409491.
[16] Hatzikirou H., Fitch W. T. S., and Herzel H. 2006. "Voice instabilities due to source-tract interactions," Acta. Acust. Acust. 92, 468–475.
[17] Miller D. G., and Schutte H. K. 2005. " 'Mixing' the registers: Glottal source or vocal tract?," Folia Phoniatr Logop 57, 278–291.
[18] Mergell P., and Herzel H. 1997. "Modeling biphonation—The role of the vocal tract," Speech Commun. 22, 141–154. doi: 10.1016/S0167-6393(97)00016-2.