

**Galunov V. I.**

**SOME PROBLEMS OF ACOUSTIC THEORY OF SPEECH  
PRODUCTION**

Saint Petersburg State University, AudiTech, Ltd  
Russia, 198904 St. Petergof, St. Petersburg, ul. Suvorovskaya, 7/2-13  
Tel/fax: 7 (812) 113-3633, e-mail: [auditech@online.ru](mailto:auditech@online.ru)

*The classical acoustic theory of speech production remains unchanged in general since the times of its creation by Helmholtz in 1870. However there is a reason to consider the whole complex of questions that yields doubts at the present times.*

Acoustic theory of speech production in its scientific layout was first formulated by Helmholtz in 1870 [1]. The main idea of this work has remained unchanged till now, and is accepted implicitly or explicitly by most of the speech scientists. It is a wide range of mathematical, methodological and technological updating that should be certainly taken in to consideration: from the work by Chiba and Kajima [2] forgotten by natural reasons, to the classical works by Fant [3] and Ungeheuer [4], and the latest works [5,6].

It seems advisable to mention two main peculiarities of Helmholtz model.

1. Speech production process consists of two independent components: the sound convolution as it is and forming of phonetic quality of the sound by convolution of resonance frequencies of articulation channel (by Helmholtz) or filtration (in modern consideration).
2. Phonetic quality of the sound is dimensioned by so called formants, or resonance frequencies of articulation canal (or the pole of transferring function of articulation filter), or maximums of the speech signal spectrum [7].

The above-mentioned peculiarities actually don't undergo any revision though it is obvious for everyone that the thing doesn't correspond to the reality. It happens because it is not clear which scientific consequences may such a revision lead to.

First of all, let's speak about the independence of sound source and articulation filter. It was demonstrated in [8] that phonetic quality of vowels is considerably formed already in larynx with no influence of articulation filter. Moreover, it is widely known from classical works that the voice source has its poles and zeros, that obviously affects the quality of the speech signal being formed. All this makes us to accept the following supposition: if there are the formants defining phonetic quality, these formants are maximums in the spectrum, but not the poles of transfer function that needs deconvolution operation, as it is assumed by classical theory.

The second question is whether the processes of speech production and speech perception are symmetric as it is assumed by classical theory. There is no doubt that it is possible to achieve certain results in phonetic quality of the sound with the help of such formants (speech of birds is a perfect example). But do namely these spectrum maximums define this quality? The first doubts arose already in the 30<sup>th</sup> after a band vocoder was created [9]. In early 60<sup>th</sup> a theory of speech legibility calculation was formed on the grounds of a large experimental material. The theory assumed a band representation of speech signal as a basis [10,11], and did not take formants into account (and it should be mentioned that the term 'formant legibility' was used to refer to the main calculation parameter in the Russian-language version of the theory). All these things had become a reason to introduce the following hypothesis (L. Varshavsky and I. Litvak): phonetic quality of sounds is defined by a certain level of power ratios in spectrum bands, and the formants (i.e. the spectrum maximums) are only an available way to achieve necessary band ratios for speech production apparatus.

On the grounds of the above-mentioned twofold ideology of principles of formation of acoustic appearance of speech signals at phonetic level, it is natural to suppose that there are several parallel functioning systems of sound-distinctive signs. It is due to existence of several different type systems of signs that provide for stableness of communication system in relation to the influence of wide diapason interference, of noises and formants organized in a certain way, regardless all above-mentioned possibilities of distortion.

The next question is: how many formants exist? Arrangement of the formants for the Russian vowels pronounced in isolated position or in syllables is known. In general, there are no notable deviations from the results obtained in Fant's classical work [3]. However it is known that the 'formant' figure is rather different in a real speech signal. One can see presence of extra maximums, splitting of maximums in the point of the formant's ideal location, and absence of classical formant maximums. The said outstanding deviations may appear as a result of influence of the 3 factors at least:

- speaker individual peculiarities;
- context;
- situational peculiarities (speaker psycho-physiologic condition, pronunciation manner in certain auditory, etc.)

If we accept the hypothesis on articulation canal having limited acoustic, resonance and filtering characteristics, but at the same time a human producing speech sounds tries to reach a certain structure of distinctive signs, one of which is a system of formants organized in a certain way, regardless of all above-mentioned possibilities of distortion of 'correct' formant structure, the zones of the said 'correct' formant structures will be seen on the histograms of distribution probability of spectrum maximums.

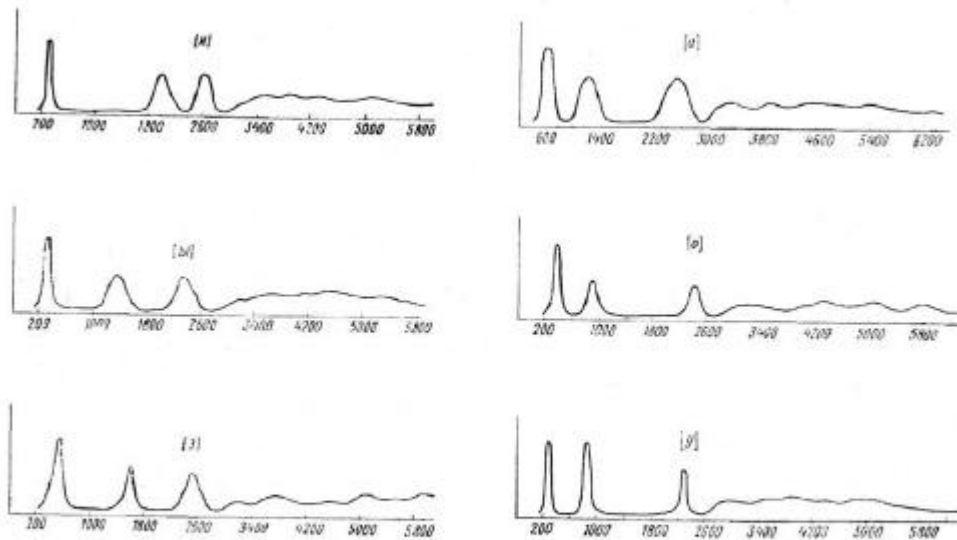
It was shown in the work [12] that for all the vowels in a low-frequency spectrum part on probability distribution of spectrum maximums appearance there are three clearly seen maximums corresponding to three major formants. In spare intervals and in high-frequency area probability distribution of maximum appearance has almost uniform motion characteristics and it's impossible to mark any formant area. At the same time the mentioned formant zones correspond to the commonly accepted in scientific practice concept of three major formants of the Russian vowels. The only thing that should be pointed out is that in continues speech for male voices there is an additional spectrum maximum that usually appears between the 2<sup>nd</sup> and the 3<sup>rd</sup> formants. What happens for female voices is that the 2<sup>nd</sup> formant usually disappears. (This phenomenon is even more typical for children's voices).

Because of the problem of the significance of the formants for perception one more problem that is not actually discussed should be pointed out. It is known that for real speech it is the merge intervals between sounds that play the main role for the formation of linguistic conception of what was said. Stationary intervals may be even exchanged one by another with no damage for sense. It's unclear how can it correspond to the explicit value of stationary intervals, with their formant structure in isolated vowel or syllable perception.

Now let's discuss possibilities of band representation of speech signal. It is obvious that a man is unable to control a large number of spectrum constituents in the process of speech production because of purely physical limitations. It is consequent to physical limitations of possibilities of voice channel [13]. If at a certain moment we measured even a large number of spectrum constituents, they may change in time only correlatively. Independent spectrum constituents may be discovered by the analysis of correlation matrices of time envelopes of spectrum constituents. It is natural to suppose that the independent constituents obtained in such a way are namely those substantial variables that define quality of the signal. In the work [12] by means of factor analysis of correlation matrices of spectrum constituents of speech material the following approximately independent spectral bands were obtained: 80-400, 400-750, 750-1350, 1350-1750, 1750-2200, 2200-2900, 2900-5000Hz. The elicited constituents correspond well to the bounds of formant bands (it is interesting that the band of 1350-1750Hz correspond to no formant of Russian vowels but it gives often the pseudo-formant for male voices (see the previous section).

Let's return back to the traditional scheme. The main assumption is that information sound wave is formed in speech tract by means of excitation of natural oscillations of sound waveguide which speech tract is assumed to be. From the open end of the waveguide the waves propagate in elastic medium and finally excite the natural oscillation of waveguide that represents speech tract. These natural frequencies are analyzed in the brain and are perceived by human as speech. From the theoretical point of view such scheme surely can exist and it has more than few successes. Yet there are moments that from our point of view poorly conform to reality. Firstly the spectra of the same

signals from different sources often differ greatly, that violates drastically the traditional scheme. Moreover it is not at all clear what happens in the case of speech with considerable change of articulatory tract (e.g. with a cigarette in a mouth). Moreover while a speech signal is transmitted through a phone channel its spectrum is distorted significantly without distortion of its meaning. This fact alone would be sufficient to throw doubt upon the adequacy of the traditional model of speech generation. There are still other doubts that can be considered emotional however. What for does the nature need such complex system? Indeed, in order to form a certain time dependence on the end of waveguide not only certain modes of waveguide must be excited but they must be excited with certain phase and amplitude ratios. And the role of large spatial deformations of speech tract in the process of speech production is absolutely unclear.



At least these questions are solved in another model that we call the modulation model of speech production (MMSP). Here conventionally two stages can be distinguished. At the first stage a sound wave appears that contains no information and plays the role of a carrier. At the second stage the carrier is modulated and this modulation contains all speech information. In this model the role of speech tract is reduced to the role of modulator, and thus the considerable spatial speech tract deformations becomes natural. The problem of strict phase-amplitude ratios between natural frequencies is solved because the carrier may be (and it is, from our point of view) of random nature. This, by the way, allows to account for the spectrum discrepancy of the same speech signals and therefore the peculiarities of transmission through phone channels. One of the possible modulation schemes is proposed in another paper [14].

## REFERENCES

1. Helmholtz H., Die Lehre von der Tonempfindungen als physiologische Grundlage für die Theorie der Musik, Braunschweig, 1870.
2. Chiba T., Kajima M., The vowel, its nature and structure, Tokyo, 1941.
3. Fant G. Acoustic Theory of Speech Production. Moscow, Nauka, 1964 (In Russian).
4. Ungeheuer G. Elemente einer akustischen Theorie der Vokalartikulation. Berlin, Springer, 1962.

5. Kent R.D. et al. (Eds) *Papers in Speech Communication: Speech production*, Ac. Soc. of America, 1991.
6. Sorokin V.N. *Theory of Speech Production*. Moscow, Radio and Svyaz, 1985 (In Russian).
7. Flanagan, J.L. *Speech Analysis, Synthesis, and Perception*, Moscow, Svyaz, 1968 (In Russian).
8. Galunov V.I., Krylov B.S., Stankevich S.A., Khantemirov R.G. *The Study of Aerodynamic Processes in Larynx*. The III<sup>rd</sup> meeting of otolaryngologists of RSFSR, 1972. (In Russian).
9. Dudley H. *The Vocoder*, Bell Labs. Record 17, 122-126, 1939.
10. Kryter K.D. *Methods for the calculation and use of the articulation index*. JASA 34, 1689-1697(1962).
11. Pokrovsky N.B. *Analysis and Measurement of Speech Legibility*. Moscow, Svyazizdat, 1962. (In Russian).
12. Galunov V.I., Garbaruk V.I. *Acoustic Theory of Speech Production and the System of Phonetic Features*. Proceedings of International Conference «100 Years of Russian Experimental Phonetics». Saint Petersburg, 2001. (In Russian).
13. Galunov V.I. *The Study of Variation of Human's Speech Behavior*, Doct. thesis, 1975. (In Russian).
14. Galunov V.I., Uvarov V.K. *Once More about Mechanism of Voice Production*. (In Russian).