Phonetic Aspects of High Level of Naturalness in Speech Synthesis

Vera Evdokimova, Pavel Skrelin, Andrey Barabanov, Karina Evgrafova

Saint Petersburg State University,

7/9 Universitetskaya nab., St. Petersburg, 199034 Russia

{postmaster,skrelin,evgrafova}@phonetics.pu.ru

{andrey.barabanov}@gmail.com

4 basic approaches to modeling and synthesis of speech

- The **concatenative** synthesis for a long time (since 80s of the 20th century) has provided the best compromise between price and quality of the synthesized speech. However, its main drawback is low naturalness of sound caused by a large number of joints between the basic elements in the compilation process and a need to modify the acoustic parameters (primarily, pitch, duration and spectral components).

As a result, HMM-based Unit-Selection synthesis has been focused on. Recently this model has been of great interest for research and speech applications. Uses large databases of recorded speech.

- **Articulatory -** refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there.

- **formant**, does not use human speech samples at runtime, created using additive synthesis and an acoustic model (physical modelling synthesis), called *rules-based synthesis*; however, many concatenative systems also have rules-based components.

- **parametric** synthesis

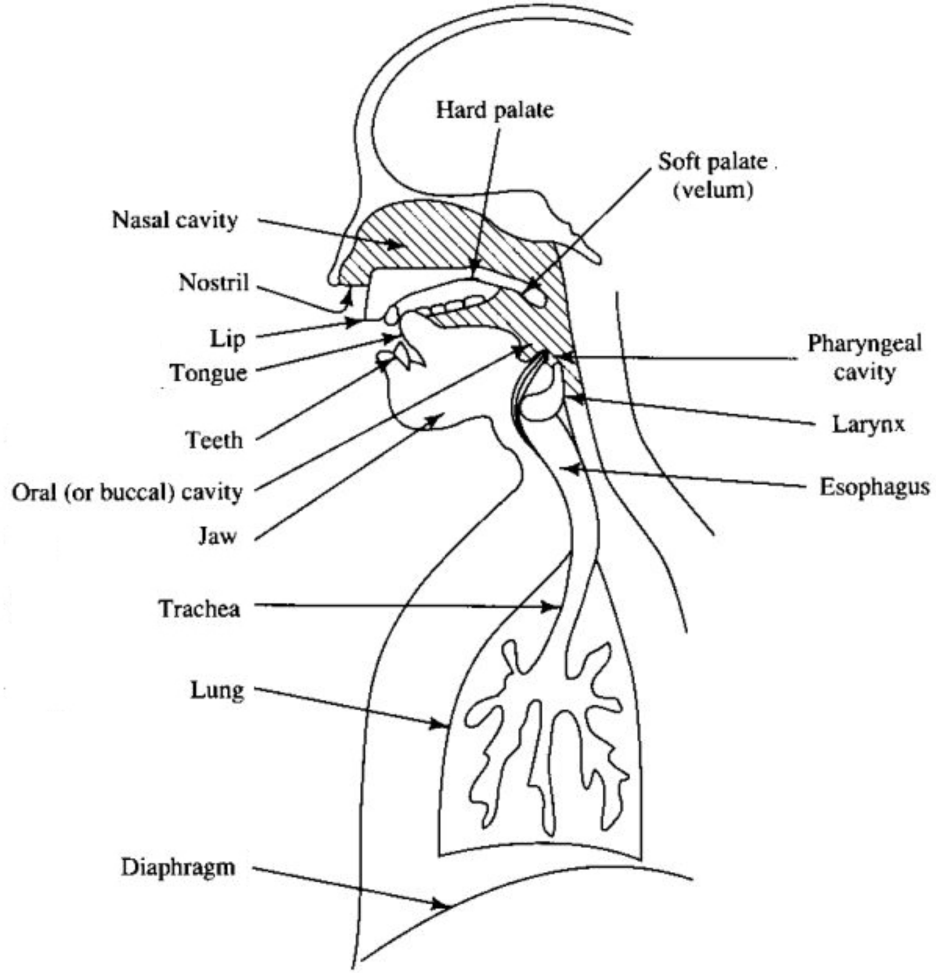
The most flexible models for speech synthesis are the **articulatory** and the **formant**. However, the synthesized speech tends to be of bad quality in terms of intelligibility and naturalness.

Aim

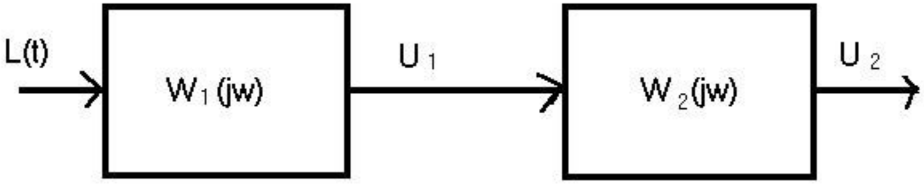
- The project is aimed at **investigating new ways to increase the naturalness of synthetic speech** on the basis of:
- 1) improvement of G. Fant‘s acoustic theory of speech production to obtain perceptually natural signals generated with the use of a parametric synthesizer;
- 2) intonation contour generation that implements necessary functions preserving naturalness and intra-speaker variability.
- The first goal supposes the comparison of acoustic parameters of a signal produced by vocal folds with the ones of the output signal. Our research is aimed at analyzing the signal of the voice source and the output speech signal to consider the non-linearity of the vocal tract system. The comparison allows calculating transfer functions for different vocal sounds in different melodic and dynamic conditions. The auditory analysis of sound streams generated by the parametric synthesizer makes it possible to obtain the maximal naturalness of artificial signals.

The traditional approach to phonetic research

- The traditional approach to phonetic research of the vocal tract assumes dividing it into two parts: the source component (vocal chords (apparatus)) and the filter component (system of articulation). The vocal apparatus consists of the vocal chords (folds), trachea, bronchi, and larynx. It is the primary source of the glottal wave. This multi-frequency acoustic signal includes the fundamental frequency and its high harmonics. The voice signal goes through the filter component – set of pharynx, nasal and oral cavities.



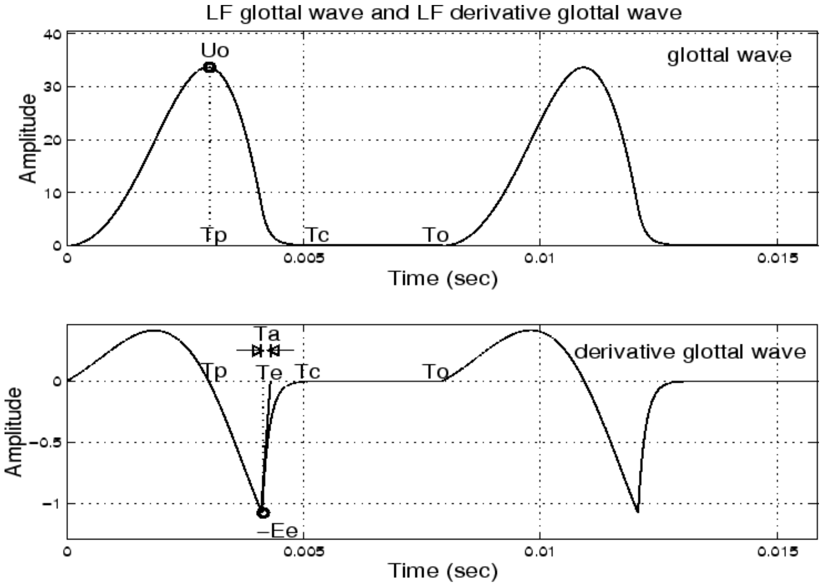
Vocal Tract Modeling



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,
U_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,
U_2(t) – speech signal.
The analysis of the signal made by MI made it possible to correct this model by adding a feedback section.

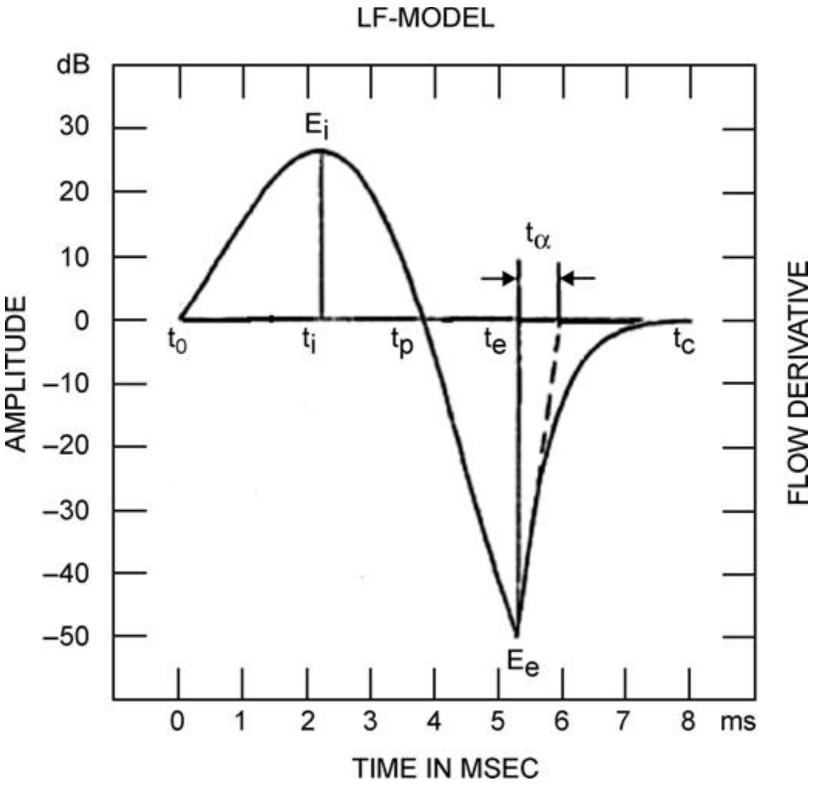
LF-model

- The LF-model of voice source was developed in 80-s by G. Fant .
- It describes 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 has shown that the voice signal can be modeled successfully by the derivative of glottal wave function.
- The glottal wave curve differs greatly from the ideal sinusoid because of the high harmonics of pitch.



LF-model

- The glottal flow is described with four different parameters. Three of these pertain to the frequency, amplitude and the exponential growth constant of a sinusoid. The fourth is the time constant of an exponential recovery.
- The choice of these four parameters provide for the production of individual voice source characteristic.
- The LF-model imitates the voice signal and works well for instrument text-to-speech synthesis system. However, it is more complicated to use it for the analysis of real speech data.



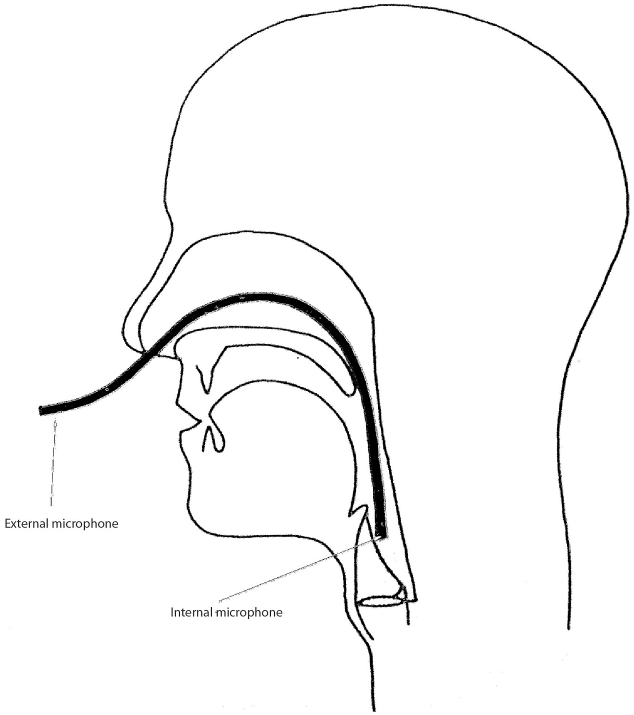
Interaction

The fact of the interaction between the two parts of the vocal tract does not make the traditional linear source–filter theory completely consistent. Obtaining the vocal fold 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.

Can we record the voice source?

Equipment

- recording studio at the Department of Phonetics, SPbSU
- Multichannel recording system Motu Traveler and WaveLab program
- sample rate of 32000 Hz and a bitrate of 16 bits
- two microphones: 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 (MI) with the use of special medical equipment.
- This procedure was performed by a phoniatrician.



Subjects

- The subjects of the experiment were 3 male and 3 female speakers. Each speaker pronounced each of the 6 Russian vowels: /a/, /e/, /i/, /ɨ/, /o/ and /u/ in different pitch modes: comfort, high, low, rising and falling. Apart from the isolated vowels the speakers were asked to read a set of words.

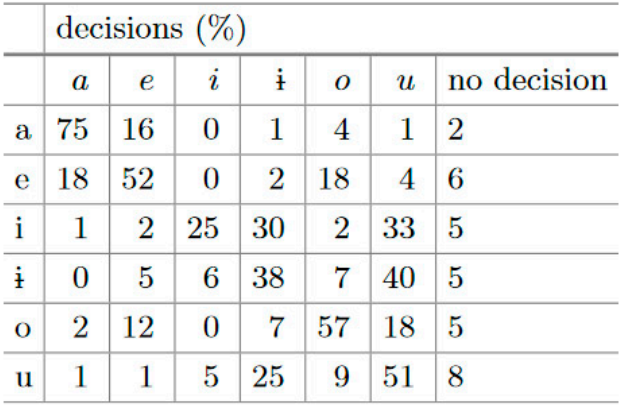
Perceptual tests

- The aim of the experiment was to find out if a voice source signal could be identified as a speech sound and which Russian vowel it could be associated with. A group of informants (23 individuals) were involved into perceptual tests. The samples were organized on a random basis. The informants were asked to assign each stimulus to one of the six Russian vowel phonemes. The questionnaire had also “no decision” option.
- The vowel sounds from the Internal microphone placed near the vocal folds were presented to informants in order to find out the way if a voice source system could be identified as a speech sound.
- The results of the perceptual tests were placed in confusion matrices which showed recognition patterns for each stimulus.

Perceptual tests

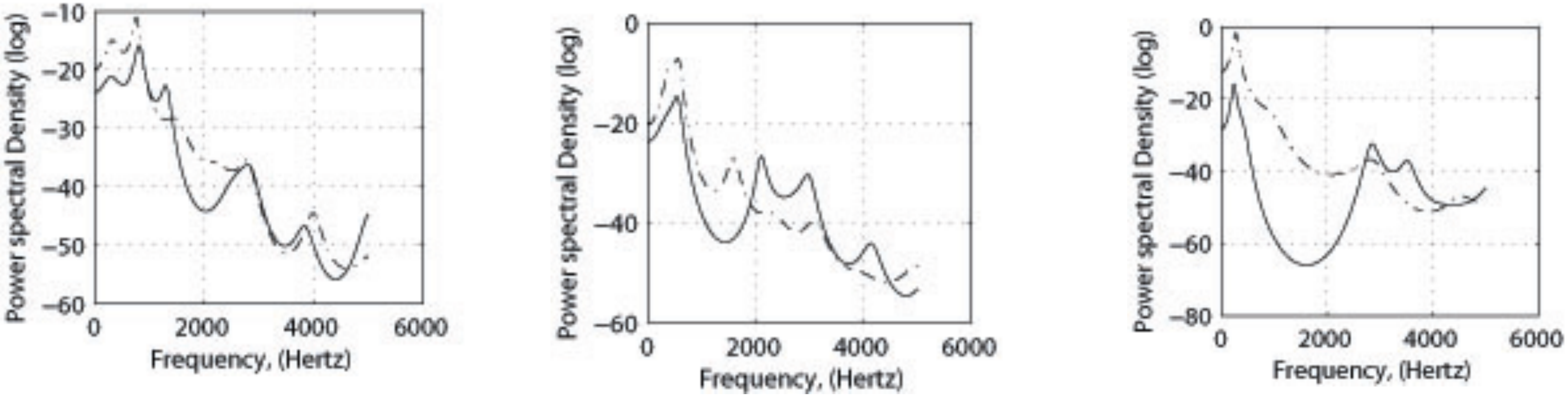
- The tests results showed that the vowel [a] stayed most intelligible and were identified correctly in most cases. The vowels [e], [o] and [u] were second intelligible (see Table 1). However, there were strong confusions of [i] and [u], [i] and [ɨ] and [u] and [ɨ]. Besides, some informants reported that all vowel types were perceived as labialized.

Table 1. Confusion matrix of vowel identification (in percentage)

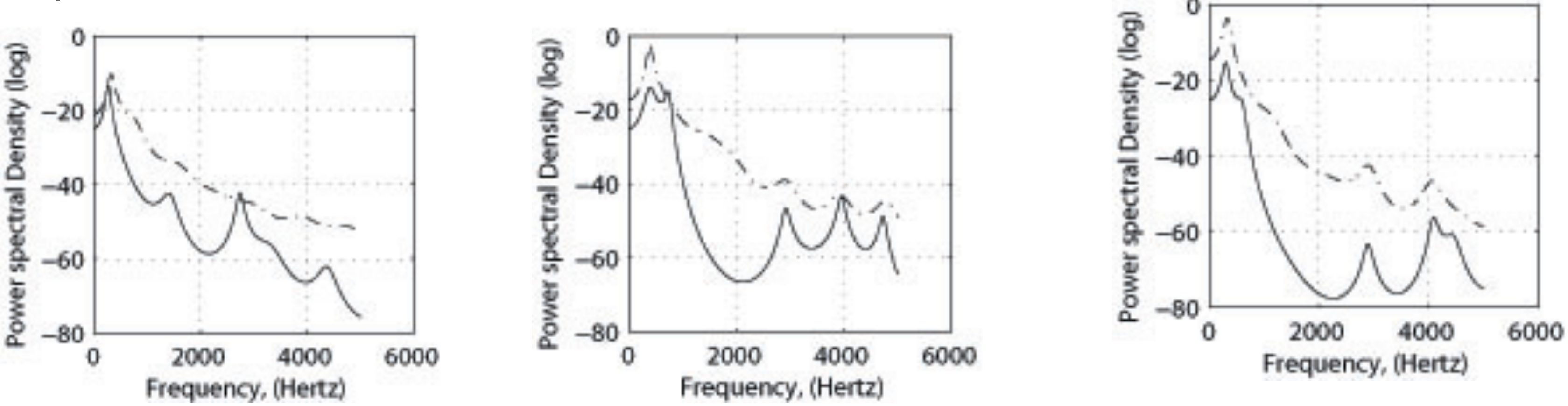


	decisions (%)						
	a	*e*	*i*	ɨ	*o*	*u*	no decision
a	75	16	0	1	4	1	2
e	18	52	0	2	18	4	6
i	1	2	25	30	2	33	5
ɨ	0	5	6	38	7	40	5
o	2	12	0	7	57	18	5
u	1	1	5	25	9	51	8

Acoustic analysis



Spectral densities for Russian vowels /a/, /e/, /i/

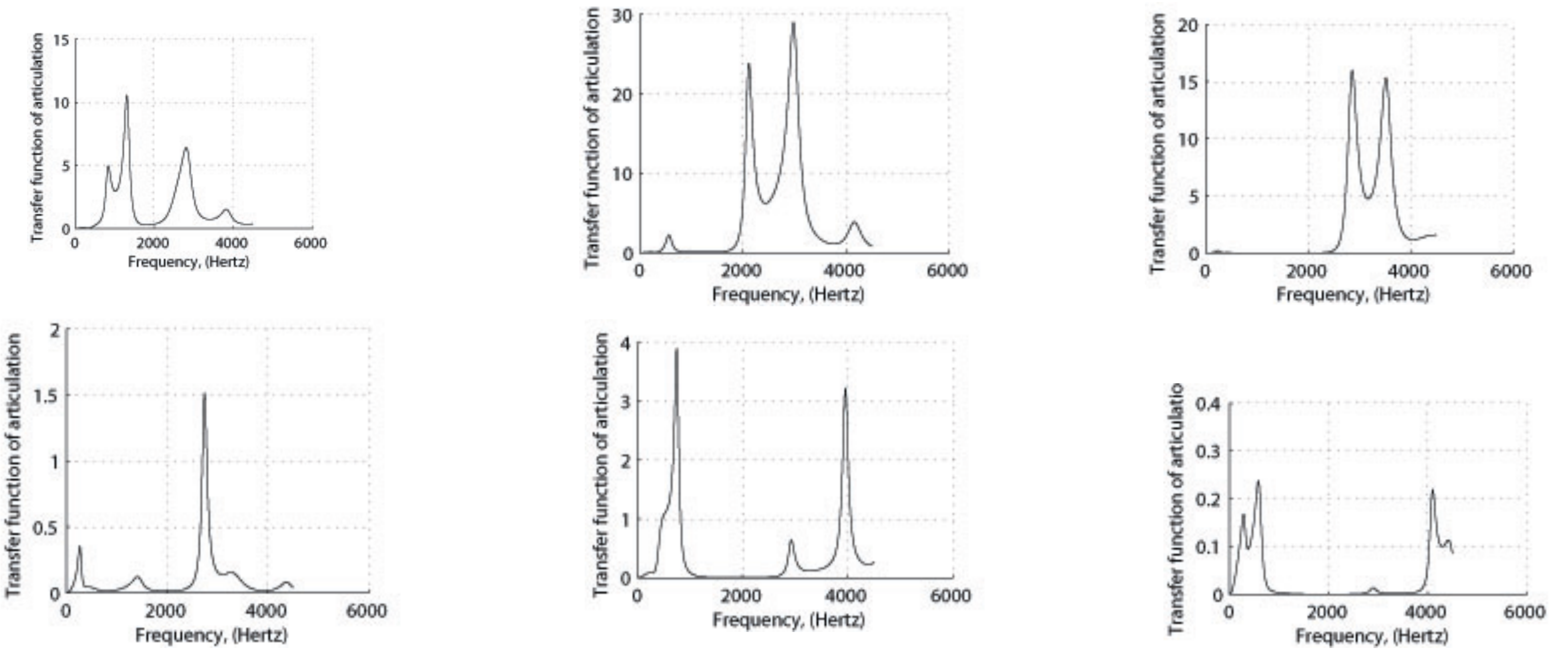


Spectral densities for Russian vowels /ɨ/, /o/, /u/

- 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 in amplitude. It can be assumed that it is caused by the reflection of the acoustic energy from the articulation system upstream. As well as this the plots show that the signals can be very different for the two microphones.
- An assumption: the acoustic analysis and the results of the perceptual experiments show that the first formant of the vowel is still recognized in the MI signal. The second and higher formants are affected more.

Transfer functions

- The next step was the discrimination and modelling of the transfer functions of the articulation. The transfer functions and the formant positions were estimated using the algorithm described in the research by Evdokimova.
- The program uses the synchronized signals from both microphones and calculates the transfer functions of the vowels.



Amplitude-frequency characteristic of the speech signal obtained from the ratio (6), filter component transfer function /W (jw)/ of the stressed vowels /a/, /e/, /i/, /ɨ/, /o/, /u/ (30 ms).

The formant structure is well-defined.

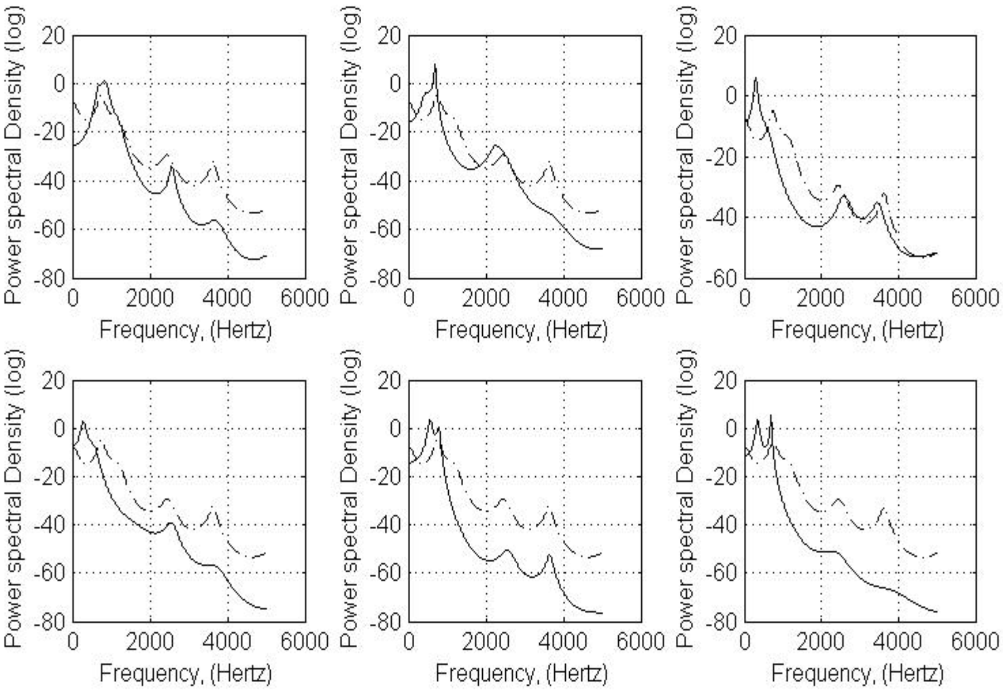
Transfer function filtering

- The obtained transfer functions of the vowels were used to generate new signals.
- The voice source signals of different vowels with different fundamental frequency characteristics were the input for these transfer functions.
- Our aim was to find out which of the following would influence the resulted signal more: the characteristics of the voice source signal or the transfer function of the articulation system.

Transfer function filtering

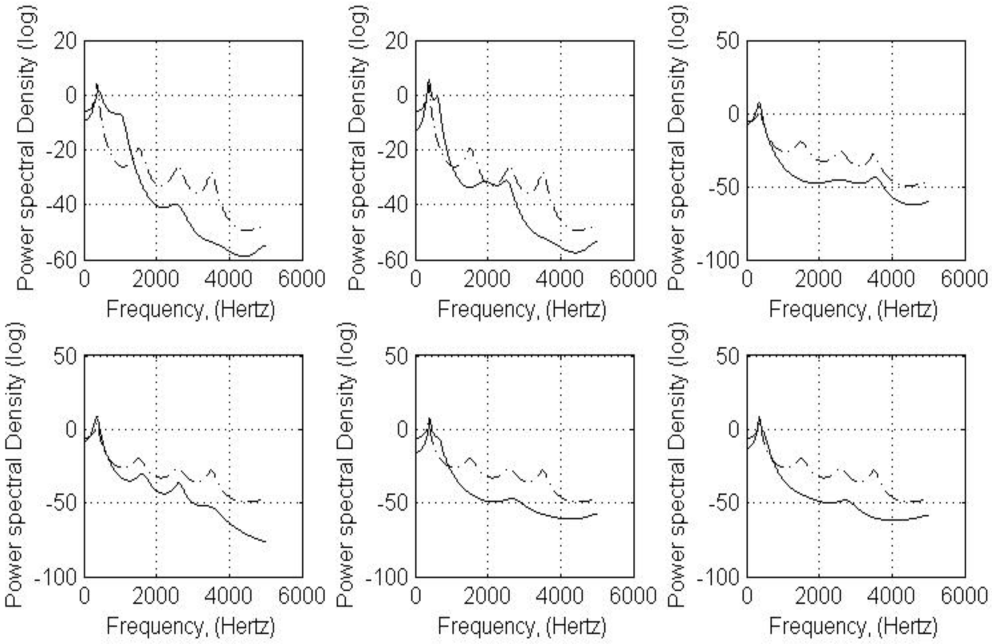
- The first experiment was the generation of the vowels using their own voice source input signal. The results showed that the produced sound had a good quality.
- The next step was the filtering of the voice source signals of higher or lower fundamental frequency. The results showed that the synthesized vowels and the input signals had similar pitch.
- The third step was to mix the transfer functions of vowel type with voice source signals of the other vowel types for the same speaker. In Fig. 2, 3 the results for two vowels are presented.

Transfer function filtering



- Spectral densities for ME signal of the input vowel /a/ (the dashed line in all six figures) and the resulted output vowels with added transfer function of articulation system for the vowels /a/, /e/, /i/, /ɨ/, /o/, /u/ (solid line). Female speaker.

Transfer function filtering



Spectral densities for ME signal of the input vowel /ɨ/ (the dashed line in all six figures) and the resulted output vowels with added transfer function of articulation system for the vowels /a/, /e/, /i/, /ɨ/, /o/, /u/ (solid line). Female speaker.

Transfer function filtering

- The final step in the filtering procedure was to mix the voice source signals and the transfer functions of the same vowel phonemes for different speakers. The MI signal of the one [a] realization was used as an input signal for the transfer function of the [a] of another speaker. The mixing of the male voices showed that the resulted vowel had the same formants as the transfer function had. However, the perceived voice quality had the characteristics of the speaker whose input voice source signal was used.

Transfer function filtering

- The results of the experiment show that it is possible to synthesize the vowel database for parametric speech synthesis.
- We should take into account the restrictions which we have. They are the type of the input vowel and the fundamental frequency and the quality of the speaker’s voice.

Conclusions

- The acoustic analysis and perception experiments allowed us to specify and improve the source-filter model. The results confirmed the fact of the interaction between the two parts of the vocal tract.
- The used approach allowed reliable automatic discrimination of the vowel formant structure by processing the speech signal.
- In experiments with mixing voice source signals and articulatory component of different vowels and speakers the resulted signal had the voice quality of the input signal.

Conclusions

- The reflected signal of the feedback section was sometimes stronger and influenced the resulting signal more in the experiments of using voice source signal of one phoneme realization with the transfer function of other vowel phoneme.
- Therefore we can conclude that the speech synthesis system should have the voice source signals for different phoneme types.
- 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.

Ongoing work

- The vowel database for 6 speakers that will be used in parametric synthesis system.

Acknowledgements

- The authors would like to thank the Saint-Petersburg State University and all our colleagues that take part in this ongoing research: Tatiana Chukaeva, Tatiana Kachkovkaya, Anton Stepikhov.
- This work is being carried out in the framework of SPbSU project n. 31.37.353.2015 and n. 6.37.349.2015 “The Phonetic Aspects of Speech Signal Synthesis with a High Degree of Naturalness”
- The authors appreciate the participation of Nikolay Shvalev, phoniatrician at the Saint-Petersburg Mariinsky Theatre, in the research.

Thank you!