The Use of Dynamic Vocal Tract Model for constructing the Formant Structure of the Vowels

Vera V. Evdokimova

Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia
postmaster@phonetics.pu.ru

Abstract

This paper discusses the new method of constructing the dynamic vocal tract model. It consists of two dynamic parts: the voice source and filter components. Each of these parts has their own dynamic features and resonant frequencies. Their interaction leads to the short-term phonetic effects. The method of obtaining frequency characteristic of the filter component by processing the real speech data is suggested. It allows constructing the formant structure of the vowels and their variations. On the example of the realization of the stressed vowel /a/ the formant structure using the new method is detected.

1. Introduction

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 [1, 2, 3]. The voice signal goes through the filter component – set of pharynx, nasal and oral cavities.

The filter component was the main object of the analysis and research in the studies of the process of speech generation for a long time. In the first physically based acoustic model of G. Fant [3] the filter component was considered as the dynamic system with the set of the resonant frequencies (the formant frequencies for vowels). The voice source signal goes to the input of this system. The voice source signal is the strongest acoustic signal in the human vocal tract. Almost all the internal organs are parts of the biomechanical oscillating system that generates the voice signal. This signal is individual and optimized by nature.

The periodic sequence of lung pressure differences in larynx is called the glottal wave [4, 5]. The frequency of these pulses corresponds to the fundamental frequency in speech signal. The shape of a glottal pulse can be similar for different people, but it can also have some differences due to size, shape, and flexibility of vocal cords.

The glottal wave generates acoustic voice signal. There is the set of poles on the plot of the spectral density of the voice source. The fundamental frequency is the lowest in frequency and the biggest in power. All other peaks are high harmonics of it (timbre frequencies). These frequencies vary slowly through the words, phrases according to the intonation contour (except tonal languages: changing of the fundamental frequency in tonal languages is important within one vowel (or syllable) for semantic differences whereas it is not important for some other languages) [6, 7, 8].

In order to provide the input excitation to the filter component in the speech synthesis model it was important to give a good description of a voice source signal. It was suggested to replace the voice source model by the description of its output signal – the glottal wave. Physiological and acoustic experiments gave an opportunity to determine the shape of a glottal pulse.

The LF-model of voice source was developed in 80-s by G. Fant [4, 5]. 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. The spectral density of voice source from the experiments was the pattern for the choice of shape of the glottal pulses. The voice source constituents were obtained from the signal using the inverse filtering. 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. 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 four parameters are interrelated by a condition of net flow gain within a fundamental period which is usually set to zero. The choice of these four parameters provide for the production of individual voice source characteristic.

The difference in quality of the speech synthesis system using the LF-model lies in its property that not only the pitch but also its high harmonics are taken into account. The basis of interference of voice and filter components is maintained in the model. The intensity of glottal flow, phoneme durations, and fundamental frequency are set as time functions for phoneme production. 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. The inverse problem should be solved in this case – to determine the LF-model parameters using the characteristics of the real speech. It is a very complicated task with lots of calculations.

2. Modeling

It can be suggested to use the method of elaboration of united human vocal tract model for the solution of the problem of analyzing the real speech data [10]. This model consists of two parts: the voice source model and the well-known model of filter component. It is suggested to extend the G.Fant’s method of elaboration the filter component model to elaborate the dynamic model of the voice source and the vocal tract on the whole. The voice source can be described as the dynamic filtered operation:

\[ L_t \rightarrow W_1(j\omega) \rightarrow U_1(t) \]

\[ U_1(t) = W_1(j\omega)L(t), \quad (1) \]

\( L(t) \) - input signal;
\( U_1(t) \) - glottal wave;
\( W_1(j\omega) \) - equivalent frequency domain transfer function of the voice source.

\( L(t) \) is supposed to be an impact of muscular and pulmonary systems (lungs) and can be set as the white noise of the given frequency bandwidth.

All the particular qualities are concentrated in the filtered operation and determined with the frequency-domain transfer function \( W_1(j\omega) \). The use of LF-model gives the basis of description of the voice source dynamic part structure.

The LF-model of the glottal wave is presupposed to consist of the fundamental frequency constituent and high harmonics. Therefore it can be suggested that \( W_1(j\omega) \) has several own resonant frequencies \( F_1, F_2, ..., F_n \). The process of glottal wave generation can be regarded as forced oscillation arising on the resonant frequencies of fundamental frequency constituent and its high harmonics’ constituents under the influence of the air flow fluctuations.

Therefore the human vocal tract can be regarded as a united dynamic system, consisting of two concatenated parts: voice source and filter component which have their own dynamic characteristics. Both parts are non-separable and interact.

\[ L(t) \rightarrow W_1(j\omega) \rightarrow U_1(t) \rightarrow W_2(j\omega) \rightarrow U_2(t) \]

\( L(t) \) – input air flow pressure from the respiratory apparatus (the lungs),
\( W_1(j\omega) \) - transfer function 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,
\( W_2(j\omega) \) - transfer function of the articulation,
\( U_2(t) \) – speech signal.
This dynamic system, consisting of two parts can be presented using the following ratios [9, 10]:

\[ U_1(t) = W_1(j\omega) L(t), \quad U_2(t) = W_2(j\omega) U_1(t), \quad (2) \]

Let us, for instance, consider the work of the vocal tract of the vowel. The \( L(t) \) signal on the input of the source component exists during the whole vowel because of the air flow from the lungs. It provides for the generation of all the frequencies in the speech signal but has no typical spectrum. The standard way of describing such signal is presenting it as a random function – the white noise of the limited frequency bandwidth [3, 11]. The limits of this band \( \omega_1 \) and \( \omega_2 \) are chosen to cover all the frequencies of the speech signal. For example in the research we can assume them to be:

\[ \omega_1 = 2\pi \times 20 \text{ s}^{-1}, \quad \omega_2 = 2\pi \times 14000 \text{ s}^{-1}, \]

The voice source with the transfer function \( W_1(j\omega) \), gains the fundamental frequency and its high harmonics. It also passes all the other frequencies, but it weakens them at the same time. Some of them are gained in the filter component having transfer function \( W_2(j\omega) \) (for example, the formants of the vowels).

Therefore there are the constituents of the source component and of the filter component in the output speech signal.

Let us find the spectral densities of the signals:

\[ S_{U_1}(\omega) = W_1(j\omega)^2 S_L(\omega), \quad (3) \]

\[ S_{U_2}(\omega) = W_2(j\omega)^2 S_{U_1}(\omega), \]

\( S_{U_2}(\omega), S_{U_1}(\omega), S_L(\omega) \) - spectral densities of the \( U_2, U_1, L \) signals.

We shall consider the procedure of detecting the parameters of the equivalent transfer functions by processing the real speech data and processing the obtained spectral densities of the \( U_2 \) signal for the vowels.

\[ S_{U_2}(\omega) = k^2 / W_1(j\omega)^2 / W_2(j\omega)^2 S_L(\omega), \quad (4) \]

\( k \) – scale factor of the experiment.

The coprocessing of several acoustic realizations helps to elaborate the methods of discrimination and modeling the transfer functions of the voice source and filter components of the vocal tract. It is important to process the speech signals that have different levels of influence of the two parts of the vocal tract. The examples of it are the processing of the 2 periods of a vowel where we can find the formant frequencies and rather long utterance where the influence of the filter component is statistically reduced but the influence of the voice source is higher.

The transfer functions of the vocal apparatus and system of articulation can be obtained through processing the experimental speech data using the ratios:

\[ / W_1(j\omega)/^2 = \frac{S_{U_20}(\omega)}{k_0^2 / W_{2II}(j\omega)^2 S_L(\omega)}, \quad (5) \]

\[ / W_2(j\omega)/^2 = \frac{k_a^2 S_{U_2a}(\omega) / W_{2II}(j\omega)^2}{k_a^2 S_{U_20}(\omega)}, \quad (6) \]

\( S_{U_20}(\omega) \) – spectral density of the output speech signal obtained by processing the long utterance,

\( S_{U_2a}(\omega) \) – spectral density of the output speech signal of the vocal tract obtained by processing several fundamental frequency periods of the vowel,

\( k_0 \) and \( k_a \) - scale gain factors for \( S_{U_20}(\omega) \) and \( S_{U_2a}(\omega) \).

\( W_{2II}(j\omega) \) – transfer function of the filter component smoothed by statistical processing of the long speech signal.

In order to get an adequate division of the voice source and filter components the described method must take into account not only the main phonetic laws but also the particular qualities of the mathematic procedures application. The use of non-parametric methods in spectral density estimation, particularly the standard procedure of the periodogram estimation leads to the irregularity of the lines in the spectral density \( S_{U_20}(\omega) \). That can lead to the mistakes in calculations.
It seems more convenient to use the parametric methods of signal processing for solving this task. In this case the spectral analysis becomes the optimization task, the search of the parameters of the model to make it as close as possible to the real speech signal [7]. The autoregression and LPC methods are used to detect the coefficients of the model. This method is known to give the good results when the spectrum of the signal has distinct peaks and high-frequency noise part.

These ratios give an opportunity to model the amplitude-frequency characteristics of the source and filter components of the vocal tract. These amplitude-frequency characteristics describe the dynamics of the system and can be used as a starting material for solving the problem of modeling of these parts of the vocal tract.

---

**Fig. 3.** Spectral density of the speech signal obtained by processing of the rather long utterance (5 minutes) of the male-voice. There is the fundamental frequency peak. The formant structure is statistically reduced.

**Fig. 4.** Spectral density of the speech signal obtained by processing of the 20 ms of the stressed vowel /a/.

**Fig. 5.** Amplitude-frequency characteristic of the speech signal obtained from the ratio (6), filter component transfer function \( |W(j\omega)|^2 \) of the stressed vowel /a/ (20 ms). The formant structure is well-defined.

**Fig. 6.** Diagram of variations of the first three formants of the stressed vowel /a/ in the word /inaunga/. This method allows describing the formant variations through the vowel.
There is no doubt that the obtained frequency characteristic does not only describe the transfer function of the filter component of the vocal tract but also contains some influence of the voice source part. There are two reasons for this.

Firstly, the voice signal in the 20 ms of one vowel is stronger than in the processing of the rather long utterance where the influence of the voice source is statistically reduced.

Secondly, the fundamental frequency of the part of the vowel is well-defined. In the processing of the rather long utterance it is statistically reduced.

3. Formant analysis

The calculations carried out show that despite some assumptions the suggested method allows to describe fully the formant structure of the vowels. The amplitude-frequency characteristics of the transfer functions of the human vocal tract parts are given as an example.

The results of the calculations show that the frequency of each of the first three formants changes during the vowel.

The set of these three frequency ranges can be the distinctive feature of the vowel. The calculations justify the phenomena that the same phoneme can be obtained by different sets of frequencies [13, 14, 15].

<table>
<thead>
<tr>
<th>F1, Hz</th>
<th>F2, Hz</th>
<th>F3, Hz</th>
<th>context</th>
</tr>
</thead>
<tbody>
<tr>
<td>360-420</td>
<td>1900-2410</td>
<td>2690-2740</td>
<td>/slabij/</td>
</tr>
<tr>
<td>470-510</td>
<td>1330-1430</td>
<td>2390-2430</td>
<td>/inaďa/</td>
</tr>
<tr>
<td>550-580</td>
<td>1060-1220</td>
<td>2070-2150</td>
<td>/prarivalis'/</td>
</tr>
<tr>
<td>580-640</td>
<td>1100-1200</td>
<td>2500-2550</td>
<td>/zahad'ašiva/</td>
</tr>
<tr>
<td>640-660</td>
<td>1210-1380</td>
<td>2300-2410</td>
<td>/naš/</td>
</tr>
</tbody>
</table>

Fig. 7. The table of the set of frequency ranges for the three first formants. Stressed vowel /a/.

4. Conclusions

1. The proposed method of describing the human vocal tract differs essentially from the well-known descriptions that use the LF-model. Firstly, it presents the voice source as an independent dynamic part with its own resonant frequencies. Secondly, the coprocessing of the acoustic realizations of one person helps to elaborate the methods of discrimination and modeling the transfer functions of the voice source and filter components of the vocal tract.

2. The proposed method gives the opportunity of automatic discrimination of the formant structure of the vowels by processing the real speech data.

3. The constructed model of the filter part of the vocal tract completely corresponds to the basic phonetic statements and can be used for solving the specific problems of speech technologies such as automatic speech recognition and high-quality speech synthesis system elaboration.

5. References

7. Sorokin V. The theory of speech production. Moscow, 1985 (in Russian)