Vera Evdokimova

Comparison of different methods of modelling the voice source in speech synthesis systems

ABSTRACT

The goal of this paper is to consider the work of the voice source as a part of the human vocal tract. This paper discusses the influences of the voice source properties on the speech signal structure. The particular qualities of speech signal defined by the voice source are marked out. The analysis of taking into account these particular qualities of the voice source was carried out for different kinds of speech synthesis systems.

Traditionally the human vocal tract is said to consist of two parts. They are the voice source and filter component. This division is quite conventional and is defined to a great extent by the functions of these parts.

The voice source is a primary oscillator of the acoustic signal and provides for the process of generating the glottal wave. Bones and cartilages of larynx take certain places while generating the sound and provide for the mechanical oscillation of the vocal cords under the action of the air flow from the lungs. The oscillation frequency of the vocal cords considerably exceeds the upper limit of the bandwidth of "brain-human vocal tract" system. The brain doesn't directly control the movements of the vocal cords. However the changes of the larynx shape are controlled by brain and change the oscillation frequency. The voice source signal is the strongest acoustic signal in the human vocal tract. Almost all internal organs (lungs, diaphragm, abdominal cavity) are parts of biomechanical oscillating system, which generates the voice signal. This signal is individual and optimized by nature [2].

Periodic sequence of lung pressure differences in larynx is called glottal wave. Frequency of these pulses corresponds to a fundamental frequency in speech signal. Changing the shape of pharynx leads to changes of glottal wave and fundamental frequency. The shape of a glottal pulse can be similar for different people, but also can have some differences due to size, shape, and flexibility of vocal cords.

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

The voice signal goes through the filter component – set of pharynx, nasal and oral cavities. Its frequency characteristic is formed under the brain control through the movements of the

articulating parts (tongue, uvula, lips and so on). There are both the constituents of frequencies of fundamental frequency and its high harmonics and constituents of frequencies of filter component of phonemes (formants for vowels, noise for some voiceless consonants) in the radiated speech signal. Their superposition leads to the short-term phonetic effects, observed in some experiments: gain or weakening of one of formants, splitting of peaks. These phenomena are randomized and hard to reproduce.

In the systems of high-quality speech synthesis two basic characteristic of voice source should be taken into account:

1. Intonation contour [3].
2. Superposition of fundamental frequency with high harmonics and frequencies of filter component.

Let's consider some of the realizations of these characteristics of voice source in different approaches to speech synthesis.

FORMANT MODEL

First efforts to imitate human speech using the electro-technical acoustic means go back to the middle of the previous century [2]. Spectral analysis of speech could make possible to get a set of harmonic constituents from the signal. Then they were reproduced. There was no difference between source and filter constituents of speech signal. It was possible to elaborate the speech synthesis system using that method but it had a bad quality. The reasons of it were not only the drawbacks of technical equipment but also not taking into account some basic phonetic laws such as prosodic structure of phrase.

CONCATENATIVE MODEL

Introduction of computers made possible the calculating of big datasets and intensified the development of concatenative speech synthesis systems. This work is often based on labeled speech databases where model predictions and database durations can be matched. The units are the segments of a real speech.

All the basic phonetic requirements (phonemic compatibility, set of allophones, coarticulation) are taken into account. The real speech units retained the speaker's voice characteristics (pitch and its high harmonics). Two factors are important for the quality of the synthesized signal. Firstly, the imitation of intonation contour and prosodic parameters requires very big databases. Secondly, qualified specialists are needed and a lot of work has to be done. It is important to find the units of the required durations, intensity and the set of constituents of frequencies. Also it is impossible to change the speaker without creating another database of

allophones. That is why the mathematical procedure of transformation of the synthesized signal is needed [13].

Sophisticated techniques have been developed to manipulate these units. The PSOLA methods are based on a pitch-synchronous overlap-add approach for concatenating waveform pieces. The frequency domain approach, FD-PSOLA, is used to modify the spectral characteristics of the signal; the time domain approach, TD-PSOLA, provides efficient solutions for real time implementation of synthesis systems [4].

The importance of PSOLA in phonetic research lies in its possibility to manipulate prosodically interesting parameters (duration, fundamental frequency and intensity) of natural speech without losing much of the original sound quality. It must be stressed that it is a nontrivial task to perform optimal PSOLA synthesis. One of key problems is to precisely define each glottal pulse in time, something that is even theoretically difficult to do for natural voices [4].

We will consider the procedure of transforming the synthesized signal suggested by Stylianou Y. in his PhD thesis as one of the most interesting and based on mathematics [16]. The order of operations is the following. Spectrum of speech signal is divided into two parts – harmonic and noise constituents. The boundary between them is chosen above the frequency of the highest formant (4000-5000 Hertz). Autocorrelation function is used to detect the fundamental frequency. The frequencies, intensities of all other harmonic constituents and the noise intensity in the signal are measured. The suggested functions give an opportunity to transform the durations of phonemes (units), fundamental frequency and intensity of signal. Harmonic signal is set as harmonic constituents; the noise part is set as a white noise of given intensity.

This procedure of Stylianou permits to change the rate, intonation contour of speech and to change the speaker. It essentially widens the capability of concatenative speech synthesis method, but it doesn't imply the whole transformation of voice source signal. Only fundamental frequency is changed. All its high harmonics are separated from it on the analysis stage and there is no difference between them and filter components (formants). They are reproduced without any transformation. It can be an essential drawback because some of the high harmonics of pitch can be rather strong in speech and their frequencies should be also transformed with the fundamental frequency.

INSTRUMENT MODEL

The development of formant synthesis led to the elaboration of the systems of instrument speech reproducing. The synthesizer input is derived from phonemic symbols instead of stored

speech units as in the concatenative synthesis. These systems of instrument synthesis have similar basic principles and structure.

The internal structure is not a model of the acoustic speech production in the vocal tract. The basic concept is the combination of sound sources and filters, describing the transfer function. The vocal tract transfer function is simulated by a sequence of second order filters in cascade while a parallel structure is used mostly for the synthesis of consonants. The Klatt model is widely used in research both for general synthesis purposes and for perceptual experiments [5, 12]. A simplified version of this system is used in many commercial products like DECtalk software of Digital [8]. Let's consider DECtalk text-to-speech system as an example. Its main properties are:

- Detailed linguistic and phonetic preliminary text processing, creating the data arrays describing the power and frequency characteristics of the signal as time functions.
- Multi-channel generating system of the speech signal.

The filter component of human vocal tract is modeled with the set of band-pass filters (which pass frequencies in the definite bandwidth). Each filter (also the formant filters) is controlled according to the predetermined program. Two types of external actions go to the input of the system: sequence of rectangular pulses and white noise. The structure of filtering system provides for the quality of the output signal. Some parts generate fundamental frequency, formants, take into account the nasal factor, and form the noise frequencies by processing the white noise which goes to the input of system. This system works as a musical instrument. Its parts are not structurally connected and there is little interference between them. The main reasons for the limited success in formant-based synthesis can be explained by incomplete phonetic knowledge. The quality of output signal is determined by the quality of phonetic processing on the preliminary stage. The internal structure is not a model of the acoustic speech production in the vocal tract. No division in harmonic constituents as belonging to the source or filter component. The standard signals substitute voice source in these systems.

ACOUSTIC MODEL

In the middle of the previous century on the basis of processing a lot of acoustic data and physiological experiments G. Fant has set up a hypothesis that human vocal tract can be interpreted as the acoustic system of given structure with the set of resonance frequencies [9].

The acoustic model consisted of two main components: voice source and filter component. The research of the first stage was mainly concentrated on the acoustic characteristics of filter component and on experimental proof of choosing this structure of a model.

This acoustic model was the first physically based model of the vocal tract. It has resulted in the appearance of the acoustic phonetics as a part of general phonetics, explaining the process of speech generation from a physiological and physical point of view.

Further work in this field made possible to get a detailed description of activity and structure of filter component and to understand their influence on its acoustic resonance characteristics. The acoustic model of filter component can be regarded as a part of speech synthesis system structure. However the voice source is as important part of vocal tract as filter component because it forms the intonation contour of speech. At first it was outside the acoustic model. However the speech synthesis problem drew the attention to the studies of voice source. The phonation process is being investigated in different countries and includes the acoustic and neurophysiologic aspects [14, 15].

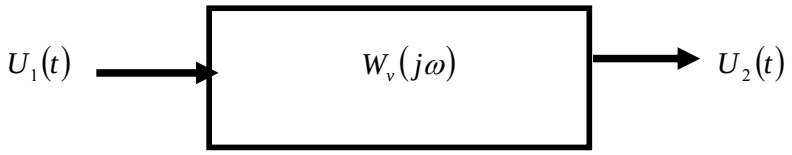
In order to provide the input excitation to the filter component in the speech synthesis model it was suggested to replace the voice source model by the description of its output signal – glottal wave. Physiological and acoustic experiments gave an opportunity to determine the shape of a glottal pulse.

The LF-model of voice source was suggested in 80-s by G. Fant [6, 10, 11]. It describes the glottal wave as a sequence of pulses of given shape. The frequency of these pulses is fundamental frequency. Their shape is similar to the experimentally measured shape of glottal pulses. The spectral density of voice source from 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 [1]. Comparing the model with the pattern has shown that the voice signal can be very well modeled by the derivative of glottal wave function. The glottal wave curve differs greatly from the ideal sinusoid because of the high harmonic of pitch. A 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 real speech. It is a very complicated task with lots of calculations.

It can be suggested to use the method of elaboration of united human vocal tract model for solution of the problem of analyzing the real speech data [7]. This model consists of two parts: voice source model and well-known model of filter component. It is suggested to extend the method of G.Fant of elaboration the filter component model to elaborate the dynamic model of voice source and the vocal tract in the whole. The voice source can be described as dynamic filtered operation:



$U_1(t)$ - input signal;

$U_2(t)$ - glottal wave;

$W_v(j\omega)$ - equivalent frequency domain transfer function of the voice source.

$U_1(t)$ is supposed to be an impact of muscular and pulmonary systems (lungs) and can be set as the white noise of a given frequency bandwidth.

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

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

Using the inverse filtering by processing the real speech data the characteristics of each of dynamic parts can be obtained.

This method differs from the G. Fant's model in the way that the dynamic parts' structure is not given a priori. Gain-frequency characteristics are detected by processing of real speech material.

The structure of these parts can be obtained using the specified mathematical procedures. The obtained model takes into account the voice source properties and perhaps can be used in speech synthesis systems. The work on elaboration of transfer functions of vocal tract is in progress.

CONCLUSIONS

1. The voice source is an oscillating system with its own resonance frequencies
2. The output signal of the voice source consists of the pitch and the constituents of its high harmonics
3. The voice source structure simplification in the speech synthesis systems leads to the low quality of synthesized signal and addition of an extra adjustment stage (possibly the greatest problem is the modeling of the filter).
4. LF-model of G. Fant gives the fullest description of the voice source signal. The use of this model in speech synthesis systems raises the quality of the synthesized speech
5. LF-model gives the basis for the elaboration the frequency transfer function of voice source through processing the real speech data.

REFERENCES

1. Akande O., Murphy P. 2005. Estimation of the vocal tract transfer function with application to glottal wave analysis. Speech Communication, p.46.
2. Bondarko L.V. Phonetics of Russian modern language, SPbSU, 1998 (in Russian)
3. Bondarenko V., Kotsubinski V., Mescheriakov R. Peculiarities of vocal generation at speech synthesis by rules. Speecom'2004, S-Pb.
4. Carlson R., Granstrom B. Speech Synthesis. The Handbook of Phonetic Sciences, Blackwell Publishers Ltd, Oxford, 1997
5. Carlson R., Granstrom B., Karlsson I. Experiments with voice modeling in speech synthesis. Speech Communication, 1991, p.10.
6. Dinther R., Veldhuis R., Kohlrausch A. 2005. Perceptual aspects of glottal-pulse parameter variations. Speech Communication, p.46.
7. Evdokimova V.V. Selection of method of human vocal tract model construction, Phonetic Lyceum, SPbSU, 2005 (in Russian)

8. Hallahan W.I. DECtalk Software: Text-to-Speech Technology and Implementation. COMPAQ DIGITAL Technical Journal. 1996
9. Fant G. Acoustic Theory of Speech Production. Moscow, 1964 (in Russian)
10. Fant G. The voice source in connected speech. Speech Communication, 1997, v. 22.
11. Fant G., Liljencrants J., Lin Q. A four-parameter model of glottal flow. STL-QPSR, № 2-3, 1985
12. Klatt D. Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. Journal of the Acoustical Society of America, 1990, v. 87, № 2.
13. Skrelin P.A. Phonetic aspects of speech technologies, SPbSU, 1999
14. Sorokin V. The theory of speech production. Moscow, 1985 (in Russian)
15. Sorokin V. Speech Synthesis, 1992. (in Russian)
16. Stylianou Y. Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification. Paris, 1996.