HIGH QUALITY SYNTHESIS OF VOWELS

Leonid Chudnovsky, Vsevolod Ageyev
Institute for Information Transmission Problem
Moscow, USSR 101447

ABSTRACT

A problem concerning synthesis of isolated Russian vowels is described. Approximation of excitation source functioning is at the centre of attention.

During vowel synthesis attention is focused on vocal tract (modes) frequency values. Excitation source is approximated by a triangular function subjected to three forms of a derivative in pitch period. That approach doesn't provide a high quality synthesis and thus causes intelligibility degradation in additive noises. Somewhat better synthesis results are achieved for a more composite time function of the vowel excitation source.

Natural sounding and intelligibility of synthesized vowels can be improved due to taking into consideration the real features of vowel excitation sources. One may get an idea of the excitation sources from vowel oscillograph traces using the inverse filtering techniques. To solve that inverse problem it is necessary to know such vocal tract parameters as a quality factor and mode frequency. A compensating method based on instant frequency measurement of a filtered speech signal has been used for mode frequency calculation. It was continued by signal frequency filtering in order to extract formant oscillations. Low-pass filters have been used for the extraction of the first formant and band-pass filters for the extraction of other formants. The cut steepness of a filter response characteristic has accounted for no less than 48 db/octave outside the transparency band. A quality factor of the extracted formant oscillations has been calculated using an analytic envelope. Algorithm 4 has been used to improve computing accuracy of a quality factor. After the extraction of formant oscillations the calculation of the quality factor Qω and the vocal tract mode frequency ωs can be performed in order to regenerate the vocal excitation source. The following equation (1) is used:

\[ P_x(t) = \frac{1}{Q_ω} P_{x0}(t) \omega_s^2 \left( \omega_s^2 \int P_{x0}(t) \, dt \right)^2 \]

(1)

The excitation source of the formant oscillation \( f_x(t) \) is related to the vocal excitation source \( f_x'(t) \) in the following way:

\[ f_x(t) = \frac{1}{Q_ω} \int f_x'(t) \left( t - t' \right) \, dt' \]

(2)

where \( L_K(t) \) is a filter pulse response for the \( K \)-th formant oscillation.

Equation (1) may be used for speech synthesis as well. Excitation sources of 5 Russian vowels "a", "3", "o", "y", "w" have been experimentally studied. The extracted excitation sources of the first formant oscillation can be conventionally divided into two groups: the first group for the sounds "a" and "3", and the second group for the sounds "o", "y", and "w". The first group of excitation sources represents \( \delta \)-successive pulses with different amplitudes with a pulse interval 4-6 ms and each pulse duration 1-2 ms. The second pulse amplitude and its delay time with respect to the first pulse are related to the quality factor and the first mode frequency in such a way that the second pulse stops its free oscillation which appeared after the first pulse. The second group of excitation source is represented either by a single pulse with 1.5-2 ms duration or by two multy- or unidirectional pulses of the same duration with the second pulse time delay 1.5-2 ms, or by the three pulses of alternating direction with the duration 1.5-2 ms and the time delay 1.5-2 ms and 1.54 ms correspondingly. Excitation source of the sound "y" has one peculiarity. The regenerate excitation sources of the first mode and the second formant oscillations are identical.

During vowel synthesis excitation pulses have been approximated by the following function:

\[ f_x(t) = \alpha(t) \frac{\int f_x'(t) \left( t - t' \right) \, dt'}{\alpha(t)} \]

(3)

The excitation source of the formant oscillation \( f_x(t) \) related to the vocal excitation source \( f_x'(t) \) is in the following way:

\[ f_x(t) = \frac{1}{Q_ω} \int f_x'(t) \left( t - t' \right) \, dt' \]

(2)

where \( \alpha(t) \) is a pulse duration; \( \alpha(t) \) is a unit function. The synthesis resulted in high intelligibility of the vowels "a", "3", "o", "y", "w" when represented by single formant oscillation. Increase in the number of formant oscillation causes intelligibility improvement. For acceptable intelligibility of the synthesized vowel "w" it should be represented by two formants. The first node excitation sources with the reduced pulse duration have been used for higher vocal tract mode frequencies (3). The duration reduction factor for the \( K \)-th mode has been selected equal to the ratio \( \alpha(t)/(\alpha(t))^K \).

The oscillograph traces of the synthesized single-formant vowels "a", "3", "o", "y", "w" and the two-formant vowel "w" are shown in Fig. 2.

Natural sounding improvement of the synthesized vowels is achieved with due regard for time variation of the excitation source parameters of each node \( f_x(t) \). Test data analysis has shown that the vowels excitation sources are subjected to different transformations, i.e. abrupt transformations with the time interval of 30-100 ms and slow period-by-period transformations. The vowels excitation sources which differ in their voice onset time with open or close vocal bands are well differentiated.

Fig. 1 shows the excitation sources oscillograph traces of the first formant \( f_{x1}(t) \), \( f_{x2}(t) \), \( f_{x3}(t) \), \( f_{x4}(t) \) and of the second formant \( f_{x5}(t) \) for the sounds "a", "3", "o", "y", "w".

Fig. 3 shows the extracted excitation sources of the first node \( f_{x1}(t) \) and \( f_{x2}(t) \) of the vowels "a" and "w" of the phonation of the vowel "a" initiates with close and of the vowel "w" with open vocal bands. Due to the extracted excitation source it has been found out that the voice onset time with open vocal bands and the cessation of phonation (Fig. 3) have the same time structure and are practically speaker independent. To achieve the vowels high quality synthesis with due regard for the source signal variation the function has been approximated with the help of the tables.
Fig. 3. The regenerated excitation sources of the first mode of the sounds "a" and "u".

The usage of excitation sources peculiarities and their relationship with vocal tract parameters gives an opportunity to achieve the high quality synthesis of vowels and speech as whole.

REFERENCES