Synthesis of some Russian utterances by rules

Pauli, S. and Derkach, M.

journal: STL-QPSR
volume: 12
number: 1
year: 1971
pages: 043-049

http://www.speech.kth.se/qpsr
II. SPEECH SYNTHESIS

A. SYNTHESIS OF SOME RUSSIAN UTTERANCES BY RULES

S. Pauli and M. Derkach*

The Purpose

The purpose of this work was to synthesize some intelligible Russian utterances with the OVE III synthesizer(1) using smoothed step commands and rules derived from a spectrographic study of symmetrical VCV-syllables pronounced by a single Russian speaker.

The OVE III Synthesizer

The control parameters are:

<table>
<thead>
<tr>
<th>Branch</th>
<th>Formants:</th>
<th>Anti-Formants:</th>
<th>Excitation:</th>
<th>Quantal Steps(dB):</th>
<th>Source:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>F1 F2 F3</td>
<td>A0</td>
<td>1</td>
<td>Voice</td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>FN</td>
<td>AN</td>
<td>8</td>
<td>Voice</td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>K1 K2</td>
<td>AK 2)</td>
<td>4</td>
<td>Noise</td>
<td></td>
</tr>
</tbody>
</table>

1) F4 and F5 are kept constant.
2) The anti-formant frequency is given by AK in dB rel. K1(+12 dB/oct) and cancels it when AK is 20 dB.
3) Single formant branch also used for non-nasal segments, e.g. the voice bar of voiced consonants.

The Synthesis Program(2)

The control parameter data are stored in the computer on a phoneme basis. Each data frame (phoneme), denoted by a one- or two-character code (e.g. S for [s], SH for [ʃ] etc.), contains specification of value and onset time (=one data point) for each of the control parameters used. It also contains information of the total phoneme duration; parameter DR**.

One or more data points/control parameter in the phoneme can be specified. Any number of control parameters can be defined in a phoneme. Thus in many cases some parameters are not specified at all. The utterance

* Guest researcher from Lvov University, USSR, February 17 - December 15, 1970.
** One DR-unit is 20 msec long.
to be synthesized is given by a string of phoneme-codes (ASHA for [afa], etc.). Thus, for each control parameter a sequence of data points is assembled.

These are interpreted as step commands; i.e. a parameter value is adopted at the time it is specified and then it is kept at this value until a new data point is given and at that very moment it jumps to the new value. Data points from one phoneme override those for the following one, if an overlap in the time direction occurs, see Fig. II-A-7a.

The set of square-waves (one sq. w. for each control parameter) is then smoothed in a digital lowpass filter prior to output to the synthesizer. The nominal time constant for the filters are:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Time constant (msec):</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>400</td>
</tr>
<tr>
<td>F1, F2, F3</td>
<td>100</td>
</tr>
<tr>
<td>FN, AK, K1, K2</td>
<td>50</td>
</tr>
<tr>
<td>Δ0, AC, ΔH, AN</td>
<td>25</td>
</tr>
</tbody>
</table>

see the figures, especially Fig. II-A-7. Because of the particular read-out speed chosen in this study the actual time constants were 25 % shorter.

Rules

The rules are of two kinds:

I. Rules inherent in the system due to the consequences of the 'smoothed square-wave philosophy'.

II. Rules relating different phoneme categories (with respect to acoustic and articulatory features) and control parameter values and timing. The rest of this article will describe the second kind of rules.

Phoneme Duration

The duration for all phonemes was chosen to 8 DR-units (160 msec). The prolongation phoneme, "::", adds 5 DR-units to the preceding vowel.

Vowels

The vowels are [u], [o], [a], [e], [i], and [\#]. The formant frequencies are given in Table II-A-1. They are derived from a spectrographic study of 300 VCV-syllables, see ref. (3). The formant data points F1, F2, F3 are specified 3 DR-units before the nominal start of the phoneme (where Δ0
gives on to 24 dB) to allow for the formants to reach their specified values for vowel phonemes in initial and isolated position.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>[u]</th>
<th>[o]</th>
<th>[a]</th>
<th>[e]</th>
<th>[i]</th>
<th>[#]</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>380</td>
<td>450</td>
<td>900</td>
<td>550</td>
<td>350</td>
<td>400</td>
</tr>
<tr>
<td>F2</td>
<td>750</td>
<td>850</td>
<td>1450</td>
<td>2000</td>
<td>2310</td>
<td>2000</td>
</tr>
<tr>
<td>F3</td>
<td>2245</td>
<td>2310</td>
<td>2400</td>
<td>2450</td>
<td>2550</td>
<td>2500</td>
</tr>
</tbody>
</table>

Table II-A-1. The vowel formant frequencies. F4 and F5 were constant at 3.5 and 4.5 kHz resp.

Consonants

The consonants were classified according to place (labials, dentals, alveolars, and velars) and manner (voiced/voiceless, fricative/stop) of articulation.

<table>
<thead>
<tr>
<th>Place</th>
<th>Manner</th>
<th>Frequency control parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fric.</td>
<td>F2</td>
</tr>
<tr>
<td></td>
<td>Stops</td>
<td>kHz</td>
</tr>
<tr>
<td>Labials</td>
<td>f  v  p b</td>
<td>.8</td>
</tr>
<tr>
<td>Dentals</td>
<td>s  z  t d</td>
<td>1.8</td>
</tr>
<tr>
<td>Alveolars</td>
<td>3</td>
<td>1.6</td>
</tr>
<tr>
<td>Velars</td>
<td>h  k  g</td>
<td>1.8</td>
</tr>
</tbody>
</table>

1) Defined by adjacent vowels

F1 is .25 kHz in fricatives and .2 kHz in stops. FN=F1 in voiced consonants.

Table II-A-2. Frequency control parameter values for consonants.

The frequency control parameters, but F1 and FN, are determined by the place of articulation. Fricatives and stops were given the values according to Table II-A-2. The timing pattern (time instants at which a control parameter changes its value) of the excitation control parameters are determined by the manner of articulation. The levels of these parameters are determined by both place and manner of articulation (Table II-A-3). In Fig. II-A-1 the timing patterns for the excitation parameters in fricatives and stops are shown.
Fig. II-A-1. TIMING PATTERNS FOR EXCITATION PARAMETERS. (smoothed step commands) A0 is shown together with AC, AH, and AN resp. The manner of articulation determines the timing pattern for the excitation parameters. The different timing patterns possible are exemplified by the dentals in [aCa]-context. AH is always switched on at the same time as AC and is kept constant throughout the phoneme. AN is switched on at the beginning of all voiced consonants and is kept constant throughout the phoneme. The actual values are given in Table II-A-3.
Table II-A-3. AC, AN, and AH levels (dB).
For A0 levels, see Fig. II-A-1.

**Place of articulation**

The target values for the frequency control-parameters have been derived from spectrograms of symmetrical VCV-syllables, where C is voiced and voiceless fricatives; [s], [z] targets have been used for the dentals etc. See Fig. II-A-2.

---

<table>
<thead>
<tr>
<th>Consonants:</th>
<th>AC</th>
<th>AN</th>
<th>AH</th>
<th>Rules:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strong</td>
<td>Weak</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fric.</td>
<td>Stops</td>
<td>Fric.</td>
<td>Stops</td>
<td></td>
</tr>
<tr>
<td>Voiceless</td>
<td>20</td>
<td>16</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>s</td>
<td>t</td>
<td>f</td>
<td>h</td>
</tr>
<tr>
<td>Voiced</td>
<td>16</td>
<td>12</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>z</td>
<td>d</td>
<td>v</td>
<td>b</td>
</tr>
</tbody>
</table>

---

Fig. II-A-2. First and second formant traces in [VzV]. The stationary part of the first vowel and the middle of the [z]-segment give the two points on each line. A mean of the [z]-points was used in the first approach as F1-F2 target value for the dental group of consonants.
Manner of Articulation

The timing pattern for AC is the same within the fricatives, merely the levels differ; Fig. II-A-1a and II-A-1b. A different timing pattern is used within all stops except for [k] and [g], where the burst is longer; thus the AC-pattern starts 1 DR-unit earlier than in the other stops. See Fig. II-A-1c and II-A-1d.

AH is always switched on at the same time as AC and is kept constant throughout the remaining part of the phoneme, (Fig. II-A-1).

There are two different timing patterns for A0: one for voiceless and one for voiced fricatives and stops, (Fig. II-A-1).

AN, finally, is switched on at the beginning of all voiced fricatives and stops and is kept constant throughout the phoneme.

General Build-Up Procedure used for the Consonants

The voiceless fricatives [f], [s], [ʃ], and [h] were the first ones to be constructed. From these the voiced counterparts (with respect to place of articulation): [v], [z], and [ʒ] were made. [h] does not have a voiced counterpart in Russian.

The voiceless stops were constructed from the corresponding voiceless fricatives by switching off AC and AH in the first part of the phoneme.

The voiced stops, finally, are made from the voiceless stops in the same way as the voiced fricatives are made from the voiceless fricatives.

Lists with synthetic VCV-words (C=fricatives and stops) have been produced. After analysis of the results from several listening tests with native listeners the formant and timing data have been changed to the values given in this text; i.e., an iterative procedure has been used to match our phoneme-data to the rule system.

Detailed Description of Fricatives and Stops

**Voiceless Fricatives (Vl.F.):** A0 and AC are switched off resp. on in two stops (24 to 12, 12 to 0 dB) at the beginning and at the end of the Vl.F. The time constants of the digital filters (see p. 44) used for smoothing A0 and AC is too short for switching these parameters on and off in only one step (from 24 to 0 dB) in fricatives, (see Fig. II-A-1a). The Vl.F. have been divided into two subgroups: strong [s], [ʃ] and weak [f], [h] with
respect to the required AC-level. For the strong group it is 8 dB higher than that for the weak, i.e.: $AC_{\text{strong/weak}} = +8$ dB. AH is kept constant during the phoneme. Fig. II-A-3 shows natural and synthetic [aCa]-words, C=V1.F.

**Voiced Fricatives (V.F.):** These are constructed from the Vl.F., see Fig. II-A-1b. The voicing is accomplished by (a) Switching A0 off later and on earlier in the V.F. than in the corresponding Vl.F. Compare Fig. II-A-1a and II-A-1b. (b) Switching on the nasal branch to provide a voice bar ($AN = 8$ dB, $FN = 250$ Hz). The tense/lax component of the V./Vl. distinction is realized by the rule: $AC(Vl.)/AC(V.) = +4$ dB (Table II-A-3). Fig. II-A-4 shows natural and synthetic [aCa]-words, C=V.F.

**Voiceless Stops (Vl.S.):** [p], [t], and [k] are constructed from the corresponding Vl.F. [f], [s], and [h] resp. The Vl.S. (Fig. II-A-1c) begin with a pause (6 DR-units long in [p] and [t] and 5 in [k]) after which the burst starts, i.e. AC and AH are switched on simultaneously. After 1 DR-unit AC is switched off but AH remains on. Fricatives are given 4 dB higher AC-level than the corresponding stops, see Table II-A-3. Fig. II-A-5 shows natural and synthetic [aCa]-words, C=Vl.S.

**Voiced Stops (V.S.):** The general Vl.F./V.F. relations hold for Vl.S./V.S. also. Thus the voicing in V.S. was achieved by adding the "nasal" voice bar. AN as in V.F. but FN = 200 Hz, see Fig. II-A-1d and Table II-A-3. Fig. II-A-6 shows natural and synthetic [aCa] words, C=V.S.

**Palatalization**

The soft-hard relation in the synthesis has been based on a study by Derkach(3), and is achieved by inserting a palatalization command, i.e. a zero-duration phoneme with [i]-formants before the consonant to be palatalized. These formant-targets which override those of the consonant, are reached in the consonant and kept there until the following vowel starts. The transitions will thus be delayed compared with a hard CV-segment, see Fig. II-A-7. Palatalization is denoted by "'" after the palatalized consonant. Spectrograms of natural and synthetic palatalized [aC'a]-words (C=dental consonant) are shown in Fig. II-A-8. No confusions about the hard/soft distinction in VCV-syllables were made in a test with 4 Russian listeners.
Fig. II-A-3. Spectrograms of natural and synthetic [aCa]-words with the voiceless fricatives [f], [s], [ʃ], and [h].
Fig. II-A-4. Spectrograms of natural and synthetic [aCa]-
words with the voiced fricatives [v], [z], and [j].
Fig. II-A-5. Spectrograms of natural and synthetic [aCa]-words with the voiceless stops [p], [t], and [k].
Fig. II-A-6. Spectrograms of natural and synthetic [aCa]-words with the voiced stops [b], [d], and [g].
HARD-SOFT DISTINCTION. Note the [i]-like formant frequency positions in the consonant and the delayed transitions in the soft segment compared to the hard one. The palatalization command is inserted before the consonant. The command has two data-points for each of F1, F2, and F3. The two arrows in Fig. II-A-7a are pointing at the two data-points of F1. For clarity the F1 transitions in Fig. II-A-7b has been drawn in dashed lines in Fig. II-A-7a.
Fig. II-A-8. Spectrograms of natural and synthetic palatalized [aC'a]-
words with the dental consonants [s], [z], [t], and [d].
Fig. II-A-9. Spectrogram of the sentence [d' e:vuʃka ka:kt' eb'a: zavu:t]
(Young girl, what is your name?)
Results

13 sentences have been synthesized. One of them is "Young girl what is your name?". In transcription [d'ɛ:v]/ka kɑːk t'ɛbɑr zavut]. Spectrogram of this sentence is shown in Fig. II-A-9.

Conclusions

The program philosophy and the rules described makes it possible to construct very compact phoneme data frames. [h] and [g] can be improved by introducing back and front variants. The palatalization may sound exaggerated to non-Ukrainian ears. This can be improved by a change in the timing of the palatalization command. At present the system design imposes restrictions on the rate of formant transitions by a preset time constant for each control parameter. This is one of the limitations which may be removed in future developments if the quality shall be improved.

Acknowledgments

We wish to thank Johan Liljencrants, the originator of the square-wave synthesis technique, whose ideas and programming made this work possible. Many valuable discussions with Gunnar Fant gave us new aspects on many problems. We are also indebted to Si Felicetti for her editorial help, to Gunnel Karlsson who made the drawings and mountings, and to other friends and colleagues in this laboratory for their cooperation.

References: