RESEARCHES ON THE FIELD OF SYNTHESIS OF THE ESTONIAN LANGUAGE

EUGEN KÜNNAP

Institute of Cybernetics
Tallinn, Estonia, USSR 200108

This paper reports the results of synthesis of the Estonian language. Constructions of synthesizers are described and the rules of synthesis are presented.

STATISTICS OF SPOKEN ESTONIAN

Every sound has his individual character and in speech process has some influence over neighbour sounds. Therefore the frequency of occurrence of phonemes, diphonemes and trigrams were investigated.
Arbitrarily choosed segments of speech were recorded, transformed into the phonetic symbols and analysed using digital computer. Analysis was made by syntagmas, i.e. by pauses in fluent speech. Selection contains 105942 phonetic symbols, which formed 19620 words and 4923 syntagmas.
In this work 31 phonemes were distinguished. In the Estonian alphabet there are 23 letters. In foreign names and loan words we can find some other letters, from which the letter f appears most frequently. It means that in written Estonian some phonemes were designated by the same letters. Consonants /l/,/t/,/n/,/s/ and /d/ can also be palatalized, which in fig.2 are marked with an apostrophe. Estonian /s/ is pronounced unvoiced, but sometimes, when it stays between vowels or after voiced consonants, vocal cords are also used. In this case /s/ is perceived as semivoiced and marked with /z/. /n/ can be palatalized and nasal. These phonemes have also the property of distinguishing between words. Usual /n/ can be in any phonetic constructions, but nasal only in /ng/ or /nk/ combinations. It is marked with two apostrophes. /b/,/d/ and /g/ are used as the indicators of short forms of /p/,/t/ and /k/. But in some cases they differ from /p/,/t/ and /k/ not only by intensity but also by spectrum and way of pronunciation. Therefore /b/,/d/ and /g/ are taken as different phonemes and conventionally named as semivoiced plosives. In this way we have phonemes: /a,b,d,d,e,f,g, h,i,j,k,l,l,m,n,n,n'',o,p,r,s,s,z,t,t,u,v, õ,ä,ö,ü/. The three most frequent phonemes are: /a/(11,61%), /e/(11,53%) and /i/(9,88%). In our material we have 99829 different diphonemes. The three most frequent are: /st/ (1,77%), /le/(1,76%) and /te/(1,60%). All in all there are 94858 trigrams, 11917 different types. The three most frequent are: /ele/(0,55%), /ist/(0,52%) and /sel/(0,49%).
Average number of phonemes in a word was 5,4 and in a syntagma 22,5.
The Estonian language is a quantitative language. There are three distinctive degrees of length, while different degrees give the word different meanings. In written text not all of the degrees are distinguished. To obtain more natural sounding synthesized speech, in some cases 5 degrees of length are used.

SYNTHESIZER WITH ANALOG CIRCUITS

The first version of terminal synthesizer consists of four oscillators, connected in parallel, pitch impulse generator, four delay circuits, four amplifiers, summator and final amplifier. The frequencies of all oscillators, durations of delay, amplitudes of formant frequencies and the time of decay of formant frequencies are controlled by means of functional generators, described below. All oscillators are excited by pitch impulse generator. To synthesize fricatives the amplified noise of diodes was used. Four bandpass filters of noise have the range from 50 Hz up to 10 kHz. By means of this synthesizer short phrases were synthesized.

HARMONIC SYNTHESIZER

The voiced phoneme consist of formants. Each formant has his frequency, equal to the frequencies of fixed harmonics of pitch. As usual a formant is composed of most intensive neighbour including 2-3 harmonics, decaying in time. To have the oscillation of fundamental frequency and his harmonics, a generator of high frequency was constructed. Dividing the oscillation by means of trigger system the desirable harmonics were obtained. Received rectangular pulses were filtered and obtained sinusoidal oscillations were used as components of synthesized phonemes. The frequency of primary oscillation was obtained multiplying the fundamental frequency with the factors 11,9, 8,7 and 5. If the fundamental frequency has the value of 100 Hz, then the primary oscillation has the value of 5,54 MHz. Such choise of factors permits us to have all the harmonics, in practice needed to synthesize all phonemes. Primary generator consists of quartz generator of 30 MHz and generator, controlled by voltage in the range

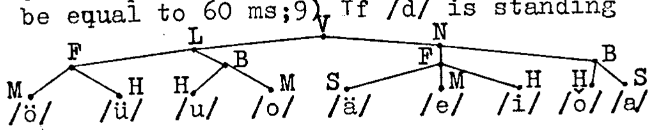
of 35 to 40 MHz. When they co-operate, the obtained beating has the range of 5 to 10 MHz and is used as primary generator. Synthesis of fricatives was made as described above.

FORMANT SYNTHESIZER WITH BANDPASS FILTERS

For the purpose to study several problems of synthesis of speech, besides synthesizer with oscillating circuits and harmonic synthesizer, a synthesizer with bandpass filters was constructed. Central frequencies of third-order analog filters were: F1 - 200÷1000 Hz, F2 - 400÷2000 Hz, F3 - 600÷3000 Hz, F4 - 1,0÷5,0 kHz, nasals F4 - 80÷400 Hz. Fixed bandwidth of filters were - 80, 120, 150, 180 and 60 Hz respectively. To synthesize fricatives the filters with central frequencies of 800÷4000 Hz and 1,2÷6,0 kHz were made. To control the central frequencies of filters the method of pulse-width modulation was used.
The transfer function of vocal tract can be realised connecting resonators in parallel or in cascade, excited by pitch impulse or noise generators. In our synthesizer both methods can be used very easily, as well as the mixed connection of filters.
To control the parameters of synthesized sounds and to have the larynx-pulse generator, which can have the output voltage in any form, corresponding generators were worked out. The form of output voltage can be easily changed in wide varieties as well as during the experiments.
In our synthesizer 12 parameters were controlled. For this purpose 12 functional generators were worked out. Each generator has the matrix of wave-form oscillation, decipher with a system of diode keys and smoothing filter. For all of generators is one common pulse generator and comparator as circular counter of 8 triggers. The number of pulses of circular counter was chosen equal to 100. Matrix of wave-form oscillation has on his surface 32 stripes of foil. Each foil is under tension taken from voltage divider in limits of 0 to -7V. Across the foils are 100 metallic wires, each of them has a sliding silver contact. In the time of each pulse from pulse generator, pulses from circular counter were given to deciphers of all functional generators at the same time and in succession they switched on voltages from dividers of all function generators to input of smoothing filters. The outputs of these filters are used to control the parameters during the synthesis. When pulse generator has the frequency within the limits of 10 to 50 Hz, the duration of speech segments can be chosen from 10 to 2 sec.
The larynx-pulse generator has the same construction as previous generators. The frequency of pulse generator is electrically controlled in limits of 8 to 25 kHz, the matrix of wave-form oscillation has 130 stripes of foil, i.e. the voltage divider has 130 levels. The fundamental frequency can be changed from 80 to 250 Hz.

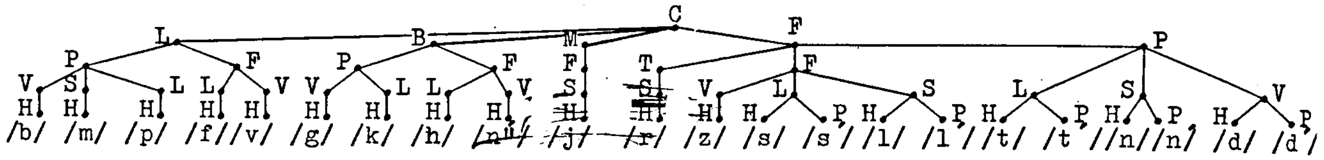
COMPUTER SYNTHESIZER

Further study of synthesis was made on the computer ES 1010. As usually, the model of vocal tract was formed by means of tunable second-order digital filters for the first three formants, fixed filter for nasals, the fourth and the fifth formants. The model consists of three branches, connected in parallel. One branch is the filter of nasals, the second consists of resonators of the third, second, first, fourth and fifth formants (fixed to 4500 Hz), connected in cascade, and the third branch consists of tunable bandpass filter for fricatives. Outputs of branches were summed up. As the source of tone the generator of triangle-form output voltage, and of noise, the generator of random numbers were used. For synthesis of nasals the branch of nasals and for synthesis of unvoiced fricatives the third branch were added to second branch. The controllable values were frequencies of pitch and first three formants, nasal and fricative formants, amplitudes of outputs of tone and noise generators, transitions, all in all 12 parameters.
To control the parameters of phonemes, they were divided by the articulary indication. In fig.1 the tree of the indication of vowels and in fig.2 the indication of consonants are shown. Every indication got his codemark and so they were stored into the memory of computer. For example phoneme /a/ has indication - VNBS, etc. The parameters of phonemes are given in table.
However, some parameters given in the table must be changed during the synthesizing process, depending on several circumstances. These changes and the rules of synthesis are as follows: 1) If a vowel stands before plosives, then the first degree of length must be equal to 40 ms; 2) Duration of vowels in diphthongs must be equal to 120 ms; 3) If a vowel stands before /f/ of /v/, then the duration of transition must be equal to 60 ms; 4) If before /r/ stands /ä/ or /ö/ and behind stands /a/, then the frequency of the first formant of /a/ must be equal to 750 Hz; 5) If before /l/ stands /ä/ and behind stands /a/, then the frequency of the first formant of /a/ must be equal to 750 Hz; 6) /b/,/d/ and /g/ in the absolute first position in word must be synthesized as /p/,/t/ and /k/ respectively; 7) /b/ in the last position in a word must be synthesized as /p/, but AN=40 units and F3==F4=800 Hz; 8) If /b/ is standing in the middle position between vowels, then the silent period before the noise burst must be equal to 60 ms; 9) If /d/ is standing



V-Vowel, L-rounded, N-unrounded, F-front, B-back, M-middle, H-high, S-low.

Fig.1



C- consonant
L- labial, B- back, M- medio-lingual, F- front
P- plosive, F- fricative, T- tremulant
V- voiced, S- sonorous, L- unvoiced; H- hard, P- palatalized

Fig.2

Table

Sym		TREE	DURATION	VOICE AMP. /AV/	NOISE AMP. /AN/	FRIC.AMP. /AF/	FORMANT FREQUENCY
0		SPACE	5	0	0	0	2500/ 1500/ 500/ 3499
1	"	SPACE	10	0	0	0	2500/ 1500/ 500/ 3499
2	A	VNBS	8	50	0	0	2500/ 1100/ 750/ 3499
3	Ä	VNFS	8	15	0	0	2280/ 1600/ 870/ 3499
4	O	VLBM	8	50	0	0	2100/ 850/ 520/ 3499
5	Ö	VLFM	8	50	0	0	2280/ 1500/ 560/ 3499
6	U	VLBH	8	50	0	0	1500/ 600/ 350/ 3499
7	Ü	VLFH	8	50	0	0	2379/ 2080/ 300/ 3499
8	E	VNFM	8	30	0	0	2843/ 2000/ 480/ 3499
9	I	VNFH	8	30	0	0	3049/ 2450/ 280/ 3499
10	Õ (Q)	VNBH	8	50	0	0	2899/ 1200/ 400/ 3499
11	B	CLPVH	8/ 2	10/ 0	0/ 20	0/ 5	10/ 660/ 400/ 2000
12	F	CLFLH	8	0	100	0	2200/ 800/ 230/ 3499
13	G	CBPVH	8/ 2	0/ 0	0/ 20	0/ 0	2100/ 1200/ 600/ 3499
14	H	CBFLH	10	0	40	0	2599/ 999/ 400/ 3499
15	J	CMFSH	6	50	0	0	3049/ 2450/ 280/ 3499
16	K	CBPLH	10/ 3	0/ 0	0/ 40	0/ 0	2100/ 1200/ 600/ 3499
17	M	CLPSH	8	50	0	0	2000/ 900/ 200/ 3499
18	P	CLPLH	10/ 3	0/ 0	0/ 0	0/ 30	2000/ 660/ 400/ 2000
19	R	CFTSH	2/ 2/ 2	30/ 4/ 30	0/ 0/ 0	0/ 0/ 0	2500/ 1500/ 500/ 3499
20	V	CLFVH	8	6	20	0	1650/ 625/ 170/ 3499
21	T	CFPLH	12/ 3	0/ 0	0/ 0	0/ 60	2599/ 1600/ 400/ 3999
22	Z,	CFFVH	8	20	20	0	2500/ 1500/ 500/ 4499
23	T'	CFPLP	16/ 1/ 1	0/ 0/ 0	0/ 0/ 0	0/ 60/0	2500/ 2000/ 600/ 3499
24	D,	CFPVH	8/ 2	0/ 0	0/ 0	0/ 10	2599/ 1700/ 400/ 3499
25	D'	CFPVP	8/ 2	0/ 0	0/ 0	0/ 20	2599/ 2000/ 600/ 3499
26	S,	CFFLH	8	0	0	3	2500/ 1500/ 500/ 4499
27	S'	CFFLP	8	0	0	7	2500/ 1500/ 500/ 4499
28	L,	CFFSH	8	30	0	0	2500/ 1400/ 400/ 3499
29	L'	CFFSP	8	30	0	0	2500/ 1700/ 400/ 3499
30	N,	CFPSH	8	50	0	0	2330/ 1800/ 200/ 3499
31	N'	CFPSP	8	50	0	0	2699/ 2100/ 310/ 3499
32	N"	CBFVH	8	50	0	0	2000/ 1250/ 290/ 3499

In the column DUR (duration) the units are in 10 ms. VA-voice amplitude, NA-noise amplitude, FA-fricative amplitude - are given in the relative units, where the unit 0 corresponds to 0 dB and 100 to 40 dB. Frequencies are expressed in Hz.

in the middle position between vowels,then the silent period before the noise burst must be equal to 80 ms and duration of vowels must be lenghtened to 120 ms; 10)/d/ in the last position in a word must be synthesized as /t/;11) If /b/ stands before /u/, then it is necessary to decrease the duration of the noise burst to 10 ms and AN to 50 units;12) /g/ in the absolute last position in a word must be synthesized as /k/, while the amplitude of noise burst must be equal to 30 units and the duration of silent period before the noise burst equal to 40 ms;13) Duration of /f/ in the first position in a word must be equal to 120 ms, and AN=60 units;14) The first degree of length of /f/ must be equal to 120 ms,the second - 180 ms and third - 240 ms; 15)/b/,/d/ and /g/ must be synthesized as /p/,/t/ and /k/ respectively,if behind them in the middle of the word stand other plosives or /s/;16)If /h/ stands before /v/, then the duration must be equal to 120 ms;17)/j/ in the first position in a word must be synthesized as /i/,only the duration must be equal to 40 ms;18) If /k/ stands before /u/,then F3 of /u/ must be equal to 1500 Hz;19)If /k/ stands in the middle position in a word,then AN=10 units 20)If /m/,/n/ or /n'/ stands in the middle position in a word,then AN=40 units;21)If /n/ stands before /a/,/o/,/u/ or /o/,then F2=1600 Hz;22)If /n/ stands behind /e/,/i/ or /ä/,then F1=250,F2=2100,F3=2500 Hz;23) If /m/ or /n/ stand between vowels in an unstressed word,then the duration must be

equal to 60 ms;24)If /m/,/n/ or /n'/ are in unstressed word,then AN=40 units;25) If /p/ is in the first position in a word, then AN=60 units;26) If/p/ is in the last position in a word, then F3=F4=800 Hz;27) If behind /l/,/m/,/n/,/r/ or /v/ in the middle of word stand /s/ or /h/, then they must be synthesized only by means of noise source;28)If plosives are in the middle position in a word, then the duration of noise burst must be equal to 5 ms;29)Duration of /v/ in the first position in a word must be equal to 120 ms,in an unstressed syllable between vowels - 40 ms; 30)If /v/ is in the first position in a word,then AN=60 units and AV=10 units;31) In compuond words between simple words must be a pause 10 ms;32)Duration of vowels of the first degree of length must be 120 ms,in an unstressed word - 80 ms,second - 180 ms,third - 240 ms, in the last position of word, when word is stressed-300ms; 33)Duration of voiced consonants of the first degree of length must be equal to 80 ms, second - 180 and third - 240 ms;34) Duration of nasals of the first degree of length must be equal to 80 ms,but if /n/ stands before vowels,then 40 ms,second for /n/ - 120 ms,for /m/ - 140 ms,third - 180 ms, in a stressed syllable - 240 ms;35)Duration of the silent period before noise burst for plosives of third degree of length - 240 ms;36)Duration of unvoiced fricatives of the first degree of length must be equal to 80 ms, second - 150 ms for /h/ and 120 ms for /s/, third - 240ms for /h/ and 280 ms for /s/.If they stand between vowels in an unstressed word,then the first degree of length must be equal to 60 ms;37)The synthesis of /n'/ and /l'/ must begin with synthesizing /i/ with the duration of 40 ms;38)The synthesis of /d'/, /t'/ and /s'/ must begin with synthesizing /i/ with the duration of 60 ms.
The durations are given for the middle rate of speech,i.e.8-10 phonemes in sec.

SYNTHESIZERS,CONTROLLED BY MICROCOMPUTERS

In the first version of terminal synthesizer,controlled by means of a microcomputer, the functional generators of the synthesizer,described above,were replaced with a microcomputer.The parameters of phonemes were stored into the memory of constants by digital keyboard.Every cell of memory has his address. To synthesize the speech signals,the contents of constant memory were fed into the memory of control parameters by means of alphabet keyboard.Both memories were connected together with a control-block of logical circuits.The task of the control block - to feed the contents of memory of constants by addresses to registers of synthesizer by commutator,using an alphabetic keyboard.The contents of all 12 registers of memory (12 controlled parameters) are fed at the same time by D/A converter to synthesizer.By means of indicator of 7 light diodes it was possible to check the contents of memories by the addresses.The table of indicator was formed of 8 diod complexes.The last version of synthesizer,controlled by means of microcomputer is more flexible and perfect.The model of vocal tract is,as described above, composed of third-order analog filters.To control the central frequencies of filters, the pulse-width modulation is also used. Central frequencies can be controlled in the range of 8 bits,but less bits are sufficient in practice.The central frequencies of the first and the third formant filters are controlled in the range of 5 bits,the second - 6 bits,filters of unvoiced fricatives and plosives 2 bits.The range of filters:F1-150+750 Hz,F2-500+2100 Hz,F3-1,5+3,2 kHz,FF-1,5+4,7 kHz,F4-3,5 kHz,F5-4,5 kHz,FN-200 Hz.The rate of transitions is controlled by 2 bits (20,40,60 and 100 ms),frequency of pitch - 3 bits (from 100 to 154 Hz),amplitudes of tone and noise generators and output amplifier - 2 bits each.ASCII - coded Estonian text is transformed into the form of discrete control signals.The microprocessor system consists of a processor unit (KP580IK80), ROM with the capacity of 6 kbyte,RAM with capacity of 2 kbyte and input-output interface.Digital control signals from the microprocessor are converted into continuous-time analog signals to control the parameters every 10 ms.

CONCLUSION

The synthesizers,described above, allowed us to research several aspects of synthesis of the Estonian language to achieve speech sounding close to natural.