EESTI NSV TEADUSTE AKADEEMIA TOIMETISED. 25. KÕIDE
ÜHISKONNATEADUSED. 1976, NR. 3

ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ТОМ 25
ОБЩЕСТВЕННЫЕ НАУКИ. 1976, № 3

https://doi.org/10.3176/hum.soc.sci.1976.3.08

G. LIIV

ON THE INTERRELATIONS BETWEEN THE SPECTRUM AND FUNDAMENTAL FREQUENCY DYNAMICS*

1. The temporal organization modeling of speech requires a thorough knowledge of the interrelations between the different timing structures defined by the distinctive parameters of its different hierarchical levels. From the viewpoint of the effectiveness and economy of the strategy of automatic speech recognition it is not only the segmentation of speech into speech units that is of great importance, but also (and, in particular) the determination of such time intervals in the latter, during which the observed speech process would with sufficient precision represent the physical correlates of the linguistic unit to be recognized. In the paper the interrelations between the formation and distribution of segments are described in the case of segmentation on the basis of the fundamental frequency (F_0) or spectral parameters (*SP*); the relations between the localizations on the time axis of the F_0 peak and the time interval of the most representative spectral composition of the syllabic nucleus are determined, and the extent of differences in spectral quality corresponding to each of the time intervals of the syllabic processes mentioned are evaluated; the possibilities are also considered of using pertinent prosodic information in automatic speech recognition.**

2. Materials, Methods and Processing.

2.1. The experimental material consists of 27 words of the ˈ$\mathrm{CV_1(:(:))CV_2}$ type. Here $\mathrm{V_1}$ denotes all the 9 stressed vowels in "short" (*Q*1), "long" (*Q*2) and "overlong" (*Q*3) degrees of quantitiy, the C's stand for alveolar consonants, and $\mathrm{V_2}$ signifies all the 4 possible vowels in an unstressed position. The words were placed at the beginning of sentences of roughly similar structure and length, not carrying the logical stress of the sentence, and the sentences were read in a random order. For the F_0 and overall intensity (*I*) measurements recordings from 6 informants (3 males (♂), 3 females (♀)) were used, in the spectral analysis from one additional female informant.

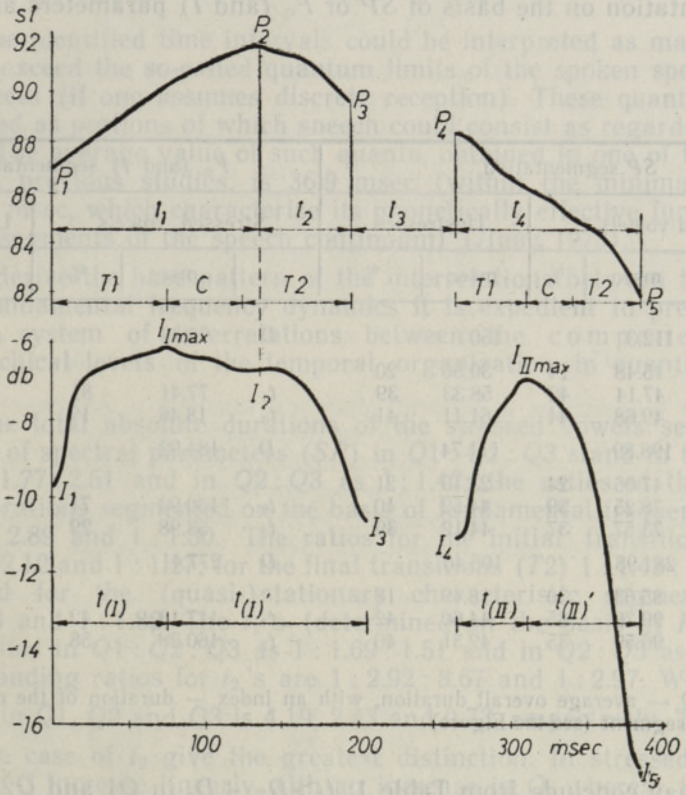
2.2. The dynamic spectrograms (*DS*) and synchronized spectral sections (the latter were recorded approximately after each 15 msec) have been obtained using a high-speed 52-channel dynamic sound spectrograph. F_0 and *I* were determined as a result of computer processing. An 8-bit A/D converter (sampling frequency of 20 kHz) was used. F_0 was obtained by

* Expanded version of paper submitted to the Eighth International Congress of Phonetic Sciences, Leeds (England), August 17—23, 1975.

** For considerations of the applicability of corresponding segmental and prosodic information in automatic recognition of speech units analyzed, see Лийв, 1974.

the shift method, and the values of F_0 and I were extracted after each 5 msec. For segmentation some harmonics were filtered out. The frequencies were transformed into absolute semitones (above 1 Hz), and the parameters were extracted after transformation (with the precision ±0.5 Hz and ±0.05 st).

2.3.1. In synchronized analysis of the *DS*'s and spectral sections the vowels were considered as consisting of three constituent segments: initial transition (*T*1), (quasi-)stationary characteristic segment (*C*), final transition (*T*2). Vowel duration values were obtained from the *DS*'s. In that connection it is significant to fix the prerequisites for the identification procedure of a quasi-stationary segment. The following prerequisites are applied for identification of *C* by spectral sections: (1) the identifiableness of a certain segment of the syllabic nucleus as a quasi-stationary one is determined by the formant frequency dynamics; *C* may thus be rendered as a certain quantum of spectral sections which predetermines its real duration in the class of accuracy of density of spectral sections; (2) in the course of *C* all formants maintain the extremum values (with some change allowed in the bandwidth). In the quasi-stationary segment, a spectral section of the most representative spectral composition (S_{CI}) is additionally localized.



The measurement points on the fundamental frequency and overall intensity curves as well as the constituent segments of syllabic nuclei are given on the basis of fundamental frequency, overall intensity and spectral parameters.

All the fundamental frequency curves in stressed vowels can be divided into a rising segment (t_1) (= the acoustic correlate of the primary stress) and a falling segment (t_2). In unstressed vowels the curves are always falling (t_4). The figure gives the normalized curves and the averaged segments (male and female informants together) of Q2 words.

2.3.2.1. In segmentation by F_0 and I parameters such initial and final values of the vowel were omitted in vowel and consonant transitions where big or opposite-direction leaps occurred in F_0 at a weak intensity level.

2.3.2.2. Our segmentation of syllabic nuclei according to F_0 parameters is based on an attempt made in a previous study to select out of the possible F_0 and I parameters those that function as the most informative ones in the quantity system of the Estonian language. For this purpose, by means of correlation analysis 7 parameters most of all correlated with the models of quantity were gradually selected from 40 F_0 input parameters and 30 I parameters (the output parameters were elementary models of quantity where 1—2—3; 1—2—2; 1—1—2; 1—2—1 corresponded to the three quantity degrees $Q1$—$Q2$—$Q3$) (the coefficients of the correlations between the parameters were also taken into consideration) — in the **1st syllable:** t_1, t_2, $t_{(I)}'$, P_2—P_3; in the **2nd syllable:** t_4, P_4—$\bar{P}_4$ (the deviations from the average); in the **1st-2nd syllables:** $(P_2$—$P_5)$: $(t_2 \div t_4)$ (see the Figure; for details, see Liiv, Remmel, 1975).

3. Results.

3.1. The differences in the formation and distribution of the segments in segmentation on the basis of SP or F_0 (and I) parameters are given in Table 1.

Table 1

SP segmentation							F_0 (and I) segmentation				
Stressed vowels				Unstressed		Stressed vowels			Unstressed	
		msec	%	msec	%		msec	%		msec
	D	112.3		150		D	95.88		t_4	151.93
$Q1$	D_c	15.48	14	30.56	20					
	D_{T1}	47.14	42	58.33	39	t_1	77.41	81		
	D_{T2}	49.68	44	61.11	41	t_2	18.46	19		
	D	198.89		111.74		D	184.92		t_4	109.59
$Q2$	D_c	47.06	24	23.10	21					
	D_{T1}	78.25	39	44.52	40	t_1	130.93	71		
	D_{T2}	73.57	37	44.12	39	t_2	53.98	29		
	D	281.98		105.40		D	277.41		t_4	105.46
$Q3$	D_c	85.63	30	18.49	18					
	D_{T1}	99.76	35	44.60	42	t_1	117.13	42		
	D_{T2}	96.59	35	42.31	40	t_2	160.28	58		

Note. D — average overall duration; with an index — duration of the corresponding constituent segment (see the Figure).

One may conclude from Table 1: $t_1 > D_{T1} + D_c$ in $Q1$ and $Q2$; in $Q_3\, t_1 <$ $< D_{T1} + D_c$. $t_2 < D_{T2}$ in both $Q1$ and $Q2$; in $Q3$ $t_2 > D_{T2}$. Consequently, the F_0 peak (P_2) in $Q1$ and $Q2$ is localized on $T2$ according to both the absolute and the relative average values, whereas in $Q3$ it is on C. A more detailed distribution can be seen when considering individual cases. In 54 cases in each Q the P_2 localization (P_{2L}) is distributed on the 3 constituent segments of the vowel as follows:

Table 2

	$Q1$	$Q2$	$Q3$
$T2$	46	38	4
C	5	15	34
$T1$	1	4	16

It follows from Table 2 that the localization of the fundamental frequency peak is shifted systematically with the increase in the quantity from the final portion of a syllabic nucleus towards the initial one.

3.2.1. In all the cases the distance of P_{2L} on the time axis was measured from the localization of the section of C with the most representative spectral composition (S_{CIL}). The respective average durations (msec;"+" marks the backwardness of P_{2L}, "−" signifies the fowardness on the time axis with respect to S_{CIL}):

Table 3

	♂	♀	♂ ♀
$Q1$	+20	+45	+33
$Q2$	+32	+52	+42
$Q3$	−18	+2	−8
$Q123$	+11	+33	+22

3.2.2. The identified time intervals could be interpreted as magnitudes that do not exceed the so-called quantum limits of the spoken speech perception process (if one assumes discrete reception). These quanta are to be interpreted as portions of which speech could consist as regards quality perception. The average value of such quanta, obtained in one of the present author's previous studies, is 36.9 msec (within the minimax limits of 20 to 100 msec, which characterize its phonetically effective functioning in different segments of the speech continuum) (Лийв, 1973).

3.3. To derive the base pattern of the interrelations between the spectrum and fundamental frequency dynamics it is expedient to present the complicated system of interrelations between the components of these hierarchical levels of the temporal organization in quantity functioning.

3.3.1. The total absolute durations of the stressed vowels segmented on the basis of spectral parameters (SP) in $Q1 : Q2 : Q3$ stand in the same ratio as 1 : 1.77 : 2.51 and in $Q2 : Q3$ as 1 : 1.42; the ratios of the corresponding durations segmented on the basis of fundamental frequency (F_0) are 1 : 1.93 : 2.89 and 1 : 1.50. The ratios for the initial transitions ($T1$) are 1 : 1.66 : 2.12 and 1 : 1.27; for the final transitions ($T2$) 1 : 1.48 : 1.94 and 1 : 1.31, and for the (quasi-)stationary characteristic segments (C) 1 : 3.04 : 5.53 and 1 : 1.82. The t_1's (determined on the basis of F_0) have the same ratios in $Q1 : Q2 : Q3$ as 1 : 1.69 : 1.51 and in $Q2 : Q3$ as 1 : 0.89. The corresponding ratios for t_2's are 1 : 2.92 : 8.67 and 1 : 2.97. We should add that $\frac{t_1}{t_2}$ in $Q1$, $Q2$ and $Q3$ is 4.19, 2.43 and 0.73, respectively. Thus, the ratios in the case of t_2 give the greatest distinction. In stressed vowels the ratios of C increase linearly with an increase in Q, whereas those for transitions increase nonlinearly and to a greater extent in $T1$.

3.3.2. The total absolute durations of unstressed vowels as a result of spectral analysis have the same ratios as 1 : 0.75 : 0.70 in Q's 1, 2, 3 of a word and in Q's 2, 3 of a word as 1 : 0.94; the ratios of the corresponding durations determined by F_0 are 1 : 0.72 : 0.69 and 1 : 0.96. The ratios for $T1$'s in Q's 1, 2 and 3 of a word are 1 : 0.76 : 0.76 and 1 : 1, for $T2$'s

1 : 0.72 : 0,69 and 1 : 0,96, and for C's 1 : 0,76 : 0.61 and 1 : 0.80. In addition we can say that $\frac{t_1}{t_4}$ in Q's 1, 2 and 3 of a word is 0.51; 1.19 and 1.11, respectively; the $\frac{t_2}{t_4}$ values are 0.12; 0.45 and 1.52, respectively. In unstressed vowels the ratios of C decrease (quasi-)linearly with an increase in Q, whereas those for transitions decrease nonlinearly ($T1$'s in $Q2$ and $Q3$ being equal) and to a greater extent in $T2$.

3.4.1. From the standpoint of possible practical applications it was essential to estimate the extent of the spectrum differences in the time intervals of syllabic processes corresponding to S_{CIL} and P_{2L}. The Euclidean distance was measured from the corresponding spectral sections according to the Expression (on the basis of F_1 and F_2):

$$d(a, b) = [\sum_{n=1}^{2} |F_{na} - F_{nb}|^2]^{1/2}$$

where d represents spectrum difference between spectral samples taken from time intervals of S_{CIL} (designated as a) and P_{2L} (designated as b). These averaged Euclidean distances are given in Table 4.

Table 4

	♂			♀		
	*Q*1	*Q*2	*Q*3	*Q*1	*Q*2	*Q*3
[a]	96	58	13	217	337	33
[e]	69	0	50	152	121	0
[i]	12	50	0	62	52	50
[o]	108	115	75	192	81	38
[u]	197	63	50	178	0	0
[ė̮]	33	56	33	267	77	67
[ä]	67	63	101	436	362	107
[ö]	98	86	0	440	142	25
[ü]	25	17	71	239	100	0

Table 4 indicates a general tendency towards a decrease in spectrum differences of S_{CIL} and P_{2L} with an increase in quantity.

3.4.2. For the reference, the Euclidean distances between all the isolated vowels were calculated (similarly on the basis of F_1 and F_2), see Table 5 (for initial data see Liiv, Remmel, 1970) (above the matrix diagonal are given the data for male informants, below the diagonal those for female ones).

Table 5

	[i]	[e]	[ä]	[ü]	[ö]	[ė̮]	[a]	[o]	[u]	
[u]	x	249	833	466	682	583	1063	1349	1616	[i]
[o]	215	x	584	289	462	849	1109	1276	1411	[e]
[a]	722	508	x	491	300	476	574	809	985	[ä]
[ė̮]	722	564	511	x	239	597	910	1034	1149	[ü]
[ö]	1281	1100	880	558	x	388	594	817	951	[ö]
[ü]	1546	1393	1179	829	299	x	378	436	562	[ė̮]
[ä]	1221	1015	581	625	507	757	x	293	493	[a]
[e]	2022	1860	1587	1300	746	485	1068	x	200	[o]
[i]	2182	2027	1764	1463	914	638	1245	178	x	[u]
	[u]	[o]	[a]	[ė̮]	[ö]	[ü]	[ä]	[e]	[i]	

3.4.3. Further, in the case of each vowel, the distance between the S_{CIL} and P_{2L} qualities (see Table 4) was compared with that between the corresponding isolated vowel and its "closest partner" (see Table 5). The analysis allows us to note that the difference in qualities between S_{CIL} and P_{2L} is considerably smaller than that between the vowel and its closest vowel type (the differences in distance are minimal or contradictory to those expected only in *Q*1 in the case of ♂ ♀ [u]; ♀ [o]; ♀ [ö], which may be due to the circumstances that the distance between the vowel types in the case of [u] and [o] is the least in the vowel system, in *Q*1 P_2 may be located in the final portion of the syllabic nucleus characterized by delabialization and, to a certain degree, probably also due to the restricted range of the material analyzed).

For the latter cases, the distances were additionally calculated between the spectrum corresponding to P_{2L} of the vowel and its "closest" vowel type (i. e. an isolated vowel), and compared to the distance between the S_{CIL} and P_{2L} qualities, for here an identification error is most probable. The former distances are as follows:

♂ [u] P_{2L}	—	[o] isol.	: 256
♀ [u] P_{2L}	—	[o] isol.	: 375
♀ [o] P_{2L}	—	[u] isol.	: 509
♀ [ö] P_{2L}	—	[ü] isol.	: 160.

Comparison reveals that the distance between the S_{CIL} and P_{2L} qualities is less than that between the P_{2L} quality and the closest vowel type * which convincingly proves the uniform base structure of the whole system.

3.5. On all the falling pitch curves of the unstressed vowels the absolute (in semitones, st) and the relative (in percent of the interval P_4—P_5) intervals were measured from the starting point of the fall to the measurement point which corresponded to the location on the time axis of the section of *C* with the most representative spectral composition. The average values:

Table 6

	♂		♀		♂ ♀	
	st	%	st	%	st	%
*Q*1	2.9	41	3.7	45	3.3	43
*Q*2	2.7	40	2.4	39	2.5	39
*Q*3	1.6	32	1.4	33	1.5	32
*Q*123	2.4	37	2.5	39	2.5	38

The most representative quality is attained in the ca $^2/_5$ interval of the whole pitch fall.

4. Conclusions.

4.1. The material analyzed in the present paper shows convincingly that fundamental frequency curves do not depend on the timing of the segments formed by the parameters of spectral dynamics, but are autonomously determined by a pattern of the interrelations between stress and distinctive quantity. Thus, the motor control system of the vocal tract and

* The only exception would be the distance between ♀ *Q*1 [ö] P_{2L} and [ü] isol., which is probably due to the two additional local factors: (1) the inter-vowels palatalized consonant-context of a word pattern (/ˈlötʲi/); (2) localization of P_2 in the immediate proximity to the boundary of syllabic nucleus and the following consonant.

the program for the fundamental frequency changes are put into effect via different control channels.

4.2. The basic results for possible practical application are as follows. In comparison with the spectral structure analysis of a speech continuum, for instance, after each 10—20 msec, the sufficiently precise determination of fundamental frequency is a considerably simpler and more reliable procedure. One may conclude from the analysis: (1) for automatic speech recognition, to identify stressed vowels it is sufficient to measure the spectrum only in a time interval corresponding to the fundamental frequency peak; (2) to identify unstressed vowels it is expedient to compile a program involving spectral measurement in a time interval corresponding to 2/5 of the complete interval of the pitch fall on the syllabic nucleus or at a distance of 50—60 msec from the beginning of the vowel.

REFERENCES

Liiv G., Remmel M., 1970. On Acoustic Distinctions in the Estonian Vowel System. — Soviet Fenno-Ugric Studies, **VI**, 1, pp. 7—23.

Лийв Г. Э., 1973. Об одном опыте оценки времени интегрирования слуховой системы человека при восприятии речи. — In: VIII Всесоюзная акустическая конференция. Рефераты докладов, I. Москва, p. 73.

Лийв Г. Э., 1974. Проблемы сегментации речи по характеристикам, выделяемым в различных частотных областях (по параметрам основной частоты и спектра). — In: Автоматическое распознавание слуховых образов (АРСО—VIII). Тезисы докладов VIII Всесоюзного семинара 16—23 сентября 1974 г., часть 2. Львов, pp. 45—48.

Liiv G., Remmel M., 1975. Estimate of the Distinctive Parameters in the Domain of Timing, Fundamental Frequency and Intensity with Implications for Modeling of a Quantitative System. — In: Fant G. (ed.), Speech Communication, Vol. 2, Speech Production and Synthesis by Rules. Proceedings of the Speech Communication Seminar, Stockholm, April 1—3, 1974. Almqvist & Wiksell International, Stockholm A Halsted Press Book, John Wiley & Sons, New York — London — Sydney — Toronto, pp. 179—185.

Liiv G., 1975. On the Interrelations between the Spectrum and Fundamental Frequency Dynamics. The Eighth International Congress of Phonetic Sciences. Leeds (England), Aug. 17—23, 1975. Abstracts of Papers. Reedprint Ltd Windsor Berkshire England: Paper No. 169.

Academy of Sciences of the Estonian SSR,
Institute of Language and Literature

Received
Feb. 12, 1976

G. LIIV

SPEKTRIKOOSTISE JA PÕHISAGEDUSE DÜNAAMIKA VASTASTIKUSEST SUHTEST

Resümee

Kõne temporaalorganisatsiooni modelleerimine nõuab erinevatel hierarhilistel tasanditel oluliste parameetritega määratavate erinevate segmentide ajastamisstruktuuride vastastikuste suhete detailset tundmist. Kirjeldatakse segmentide formeerumise ja distributsiooni vastastikuseid suhteid kõne segmenteerimisel spektraalparameetrite või põhisageduse põhjal; määratakse põhisageduse maksimumi ja silbituuma kõige representatiivsema spektrikoostisega (formandisageduste sihtväärtuste esinemisega) ajaintervalli ajateljel paiknemise vahekordi ja hinnatakse silbiprotsesside kummalegi nimetatud ajaintervallile vastava spektraalse kvaliteedi erinevuste määra; samuti käsitletakse vastava prosoodilise informatsiooni kasutamise võimalusi kõne masintajumisel.

Analüüsitud materjal osutab veenvalt, et põhisageduse kontuurid ei sõltu spektraalse dünaamika parameetrite põhjal formeeruvate segmentide ajastusest, vaid kumbagi määravad autonoomselt sõnarõhk ja distinktiivne kvantiteet. Järelikult realiseeritakse kõnetrakti motoorne juhtimissüsteem ja põhisageduse liikumise programm erinevate kanalite kaudu.

Eesti NSV Teaduste Akadeemia
Keele ja Kirjanduse Instituut

Toimetusse saabunud
12. II 1976

Г. ЛИЙВ

О ВЗАИМООТНОШЕНИЯХ МЕЖДУ ДИНАМИКОЙ СПЕКТРА И ДИНАМИКОЙ ОСНОВНОЙ ЧАСТОТЫ

Резюме

Моделирование временно́й организации речи предполагает основательное знание взаимоотношений между временны́ми структурами, определяемыми основными параметрами ее разных иерархических уровней. Описываются особенности формирования и дистрибуции сегментов речи, выделенных по параметрам основной частоты или спектра; определяются соотношения между локализациями по оси времени пика основной частоты и временно́го интервала с наиболее репрезентативным спектральным составом слогового ядра, содержащим целевые значения формантных частот, и оценивается мера различия спектрального качества в указанных интервалах времени слоговых процессов; кроме того, рассматриваются возможности применения соответствующей просодической информации для автоматического распознавания речи. Материал, проанализированный в работе, убедительно показывает, что контуры основной частоты не зависят от временно́й картины сегментов, формирующихся на основе параметров спектральной динамики, а определяются автономно взаимоотношениями ударения и долготы. Следовательно, система моторного управления голосовым трактом и программа изменения основной частоты реализуются различными каналами управления.

Институт языка и литературы
Академии наук Эстонской ССР

Поступила в редакцию
12/II 1976