Analysis of Intonation Dynamics in Macedonian for the Purpose of Text to Speech Synthesis

Branislav Gerazov, *Graduate Student Member, IEEE*, and Zoran Ivanovski, *Member, IEEE*

***Abstract* — The paper shows the results of an analysis of the intonation dynamics in spoken Macedonian. In total 7 native speakers were analyzed, with 255 recorded intonation phrases. Histograms of the respective pitch distributions were generated and analyzed statistically. The results are to be used to improve the intonation module for the text-to-speech (TTS) system "Speak Macedonian".**

***Keywords* — intonation, pitch range, speech, text-to-speech.**

I. Introduction

INTONATION, defined as the movement of pitch in time in spoken language, is one of the most significant building blocks of prosody, [1]. It serves to carry information about the discourse function, saliency, and speaker attitude and emotion. More importantly, the lack of adequate intonation can sound strange and make the speech hard to follow. This makes intonation modules a crucial building block of text-to-speech (TTS) synthesis systems. In order to build a high-quality module, a thorough intonation analysis of the language to be synthesized is required. Such an analysis has yet to be done for the Macedonian language, and only vague findings can be found in literature, [2], [3].

Being in the final stages of development of the TTS system “Speak Macedonian”, [4], the authors have undertaken such an analysis of the intonation in spoken Macedonian. Our first results, [5], describe intonation patterns in the five types of intonation phrases discerned by punctuation: declaration starts, intermediates and ends, questions and exclamations. In continuation, in this paper we analyze the intonation dynamics for the different speakers, as well as its variance across the discourse function contexts. The presented results are of significant value to the improvement of the intonation module used in our TTS system.

II. Database and Methodology

For our analysis we expanded the dataset used in [5] with recordings of an additional 7th male speaker, BG. This increased the total number of analyzed intonation phrases (IPs) to 255, and the total recording time to 9min 34s. The spread of the recording material across the speakers and intonation phrase types is given in Table 1.

Branislav Gerazov and Zoran Ivanovski are with the Faculty of Electrical Engineering and Information Technologies, Ruger Boskovik, PO Box 574 - Skopje, Macedonia (tel.: +389 2 3099 191; fax: +3892 3064262; e-mail: {gerazov@feit.ukim.edu.mk, mars@feit.ukim.edu.mk})

TABLE 1: RECORDED INTONATION PHRASES (IPS) USED IN THE ANALYSIS OF INTONATION DYNAMICS IN MACEDONIAN.

Speaker		Material length		Declarations			Quest-ions	Exclama-tions
Code	*Sex*	*Time*	*IPs*	*Start IPs*	*Interm. IPs*	*End IPs*		
VF	*m*	1m 25s	35	10	13	12	/	/
MB	*m*	1m 59s	51	14	19	18	/	/
BG	*m*	1m 16s	47	8	1	12	14	12
MV	*f*	1m 08s	26	8	11	7	/	/
ZK	*f*	1m 56s	39	12	13	14	/	/
TA	*f*	1m 33s	50	11	9	13	10	7
LU	*f*	0m 15s	7	/	/	/	7	/
Total		9m 34s	255	63	66	76	31	19

The recordings were analyzed with a variation on the classic autocorrelation pitch extraction algorithm given in [6]. The extracted pitch contours were then grouped according to speaker and discourse function. Declarations formed a single IP group, and were not divided into three different types of IPs. For each of the groups, histograms of the pitch samples were generated. In addition, cumulative histograms were also generated for speakers having different groups of IPs (BG and TA).

The histograms were analyzed using standard statistical analysis tools. The first four moments of the pitch distributions were calculated: the mean, the standard deviation, the skewness, and the kurtosis. In addition to the mean, the median and the mode were also calculated. While the mean is the average pitch, the median is the pitch value that separates the lower half from the upper half of the pitch samples and the mode is the most common pitch value in the analyzed set. These never have the same value.

The second moment was calculated through the standard deviation σ and not the variance, because it gives a better picture of the spread of the distribution in Hz. A width of $+/-\ \sigma$ around the mean encompasses 68,3 % of all the samples, and a width of $+/-\ 2\sigma$ around the mean encompasses 95,4 % of them. A width of 3σ in turn, covers 99,7 % of the samples in the distribution. The standard deviation was calculated using Eq. 1, where N is the number of pitch samples analyzed, x_i is one pitch sample, and μ is the mean of the set. The expression uses Bessel's correction (uses $N-1$ instead of N for averaging). This holds when the analyzed set is a part of a larger population, which is true in this analysis, even though the correction is minor (<0,1 Hz for σ), because of the large number of samples used.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (1)$$

The third and fourth moments are given in their normalized form, i.e. through skewness and kurtosis. The generalized form of the expression for calculating the normalized *n*-th moment of a population sample is given in Eq. 2. Skewness measures the asymmetry of a distribution around its mean. Positive skew means a longer tail to the right of the mean and the bulk of the values (with the median) to its left. Negative skew means the exact opposite. The kurtosis measures the sharpness of the peak of the distribution and the strength of its tails. Kurtosis larger than 3 (a leptokurtic distribution) means a thinner peak around the mean and stronger tails, while kurtosis smaller than 3 (a platykurtic distribution) means a wider peak around the mean and thinner tails, where 3 is the kurtosis of the normal distribution.

$$\mu_n = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^n}{\sigma^n} \qquad (2)$$

III. Results

This section presents the generated histograms and the results from their analysis. In the first part we discuss the variation of intonation dynamics across the 7 speakers. The second part analyzes the change in intonation dynamics related to discourse function in two of the speakers.

A. Intonation dynamics across speakers

The histograms of the pitch contours for the 7 speakers are given in Figs. 1 and 2. The red line gives the average pitch, while the magenta and cyan lines give the σ and 2σ ranges. The mode is the highest peak in the distributions.

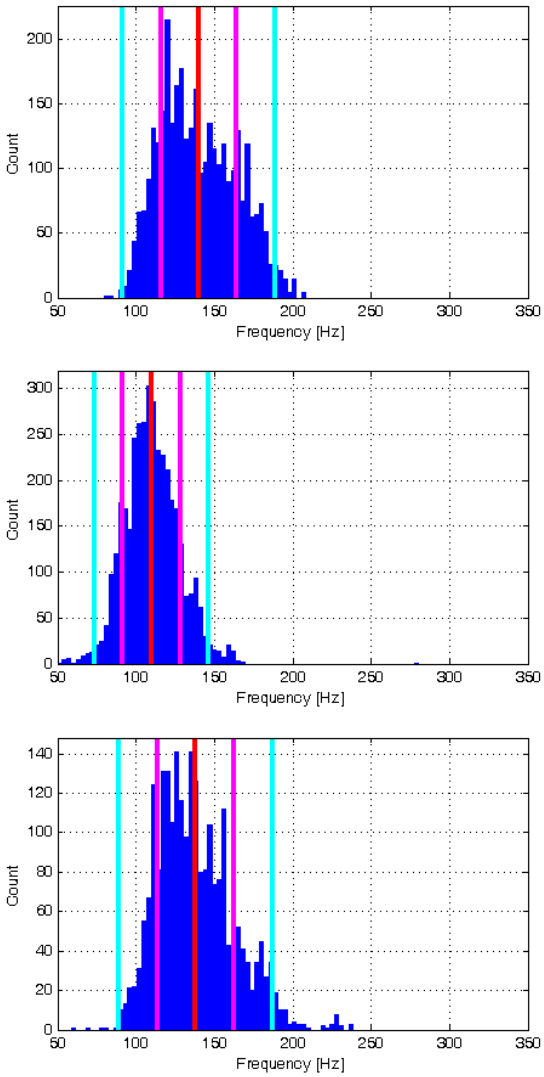

Fig. 1. Histograms for the male speakers (*top to bottom*): VF, MB, and BG.

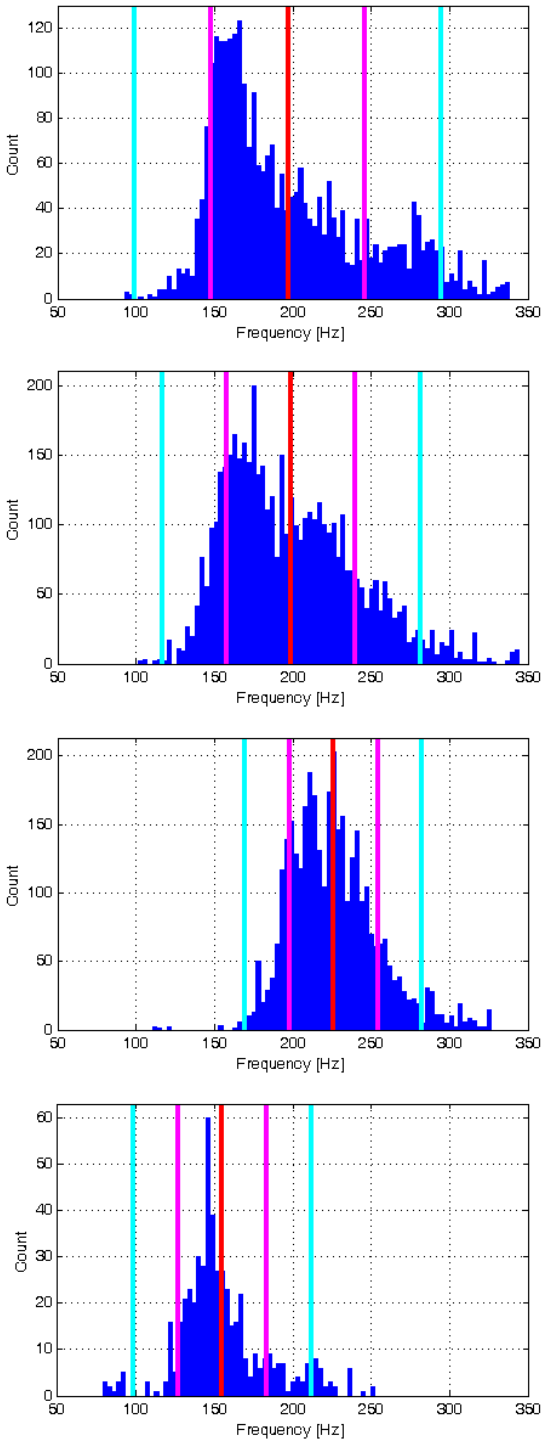

Fig. 2. Histograms for the female speakers (*top to bottom*): MV, ZK, TA, and LU.

The results from the statistical analysis of the histogram data are shown in Tables 2 and 3. Table 2 shows the statistical parameters extracted from the male speakers, and Table 3 from the female speakers. The tables also show the average value for each of the parameters.

TABLE 2: STATISTICAL PARAMETERS OF THE MALE SPEAKERS' PITCH DISTRIBUTIONS.

Variable	Speaker			
	VF	MB	BG	Average
mean	136,97	109,96	137,82	128,25
median	136,96	109,16	134,86	126,99
mode	165,17	110,53	135,69	137,13
std	24,22	18,32	24,55	22,36
std x2	48,43	36,65	49,11	44,73
std x3	72,65	54,97	73,66	67,09
skew	0,30	-0,59	0,70	0,14
kurtosis	2,25	15,50	3,65	7,13
Samples	3353	3794	2408	
	total samples		9555	

TABLE 3: STATISTICAL PARAMETERS OF THE FEMALE SPEAKERS' PITCH DISTRIBUTIONS.

Variable	Speaker				
	MV	ZK	TA	LU	Average
mean	196,62	198,56	225,87	155,46	194,13
median	181,48	191,74	223,86	148,99	186,52
mode	164,55	173,62	229,69	147,99	178,96
std	49,02	41,17	28,21	28,41	36,70
std x2	98,04	82,34	56,42	56,82	73,40
std x3	147,05	123,52	84,62	85,22	110,10
skew	0,84	0,70	0,69	0,71	0,74
kurtosis	2,79	3,17	4,04	4,15	3,54
Samples	2761	4661	3533	526	
		total samples		11481	

From the presented histograms and statistical parameters, a clear difference in the average pitch can be seen between the male and female speaker groups, as is expected. Also, the standard deviation from the average pitch is greater in women than in men. This means that women have, or use, a wider pitch range when speaking, leading to more dynamic pitch contours and more pronounced intonation.

From the histograms and skewness factor we can note that almost all of the pitch distributions are skewed to the left of the mean pitch, the mode being almost always on its left side. This means that speakers favor going further high than low from the mode pitch in shorter time intervals compared to the time they spend around it.

The kurtosis shows that most of the speakers' pitch distributions are leptokurtic (kurtosis larger than 3), having narrower peaks at the mean and larger tails than a normal distribution would. Only two of the speakers (VF and MV) have platykurtic pitch distributions with kurtosis smaller than 3.

B. Intonation dynamics across discourse function

Histograms from the pitch distributions for three different discourse functions (declarations, questions and exclamations), for two of the speakers (BG and TA), are presented in Figs. 3 and 4. The statistical analysis results for these distributions are given in Table 4.

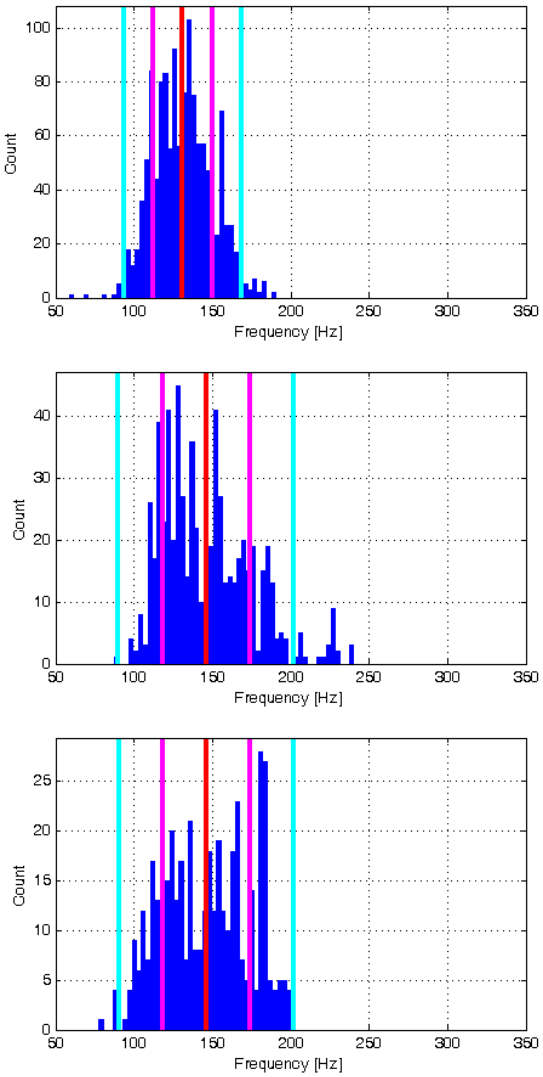

Fig. 3. Histograms for speaker BG's pitch distributions for: declarations (*top*), questions (*middle*), and exclamations (*bottom*)

TABLE 4: STATISTICAL PARAMETERS OF THE PITCH DISTRIBUTIONS FOR DIFFERENT DISCOURSE FUNCTIONS FOR SPEAKERS BG AND TA.

Variable	Speaker					
	BG			TA		
	Dec.	Quest.	Excl.	Dec.	Quest.	Excl.
mean	130,96	146,17	145,61	224,40	231,37	226,73
median	130,86	142,49	146,27	220,50	230,89	224,43
mode	135,69	146,03	180,74	211,00	232,11	223,86
std	18,80	27,95	27,80	27,78	31,39	22,72
std x2	37,61	55,89	55,60	55,56	62,77	45,45
std x3	56,41	83,84	83,39	83,34	94,16	68,17
skew	0,19	0,81	0,02	0,62	0,87	0,42
kurtosis	2,81	3,50	1,92	3,83	4,11	3,15
Samples	1306	662	440	2588	645	300

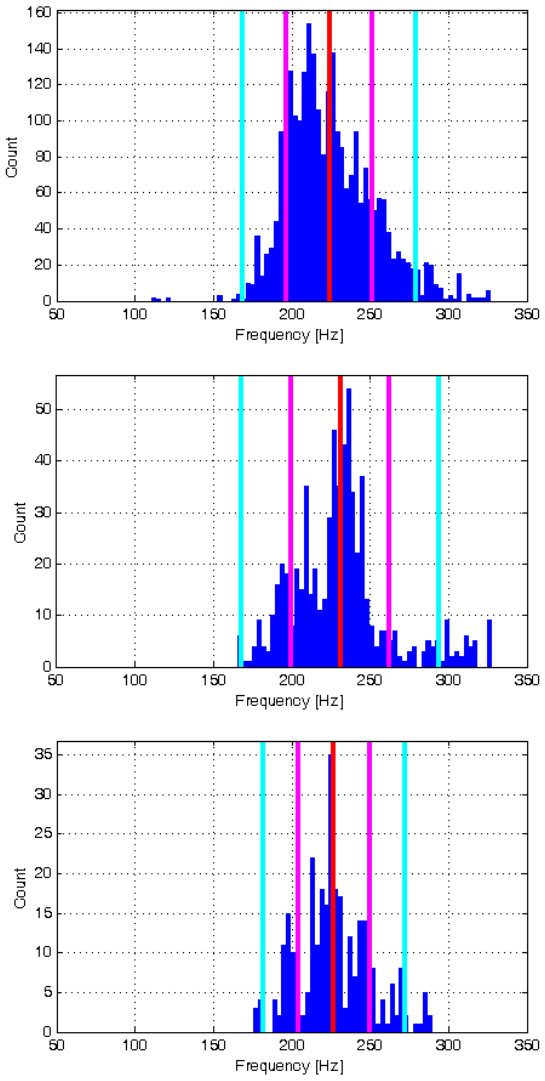

Fig. 4. Histograms for speaker TA's pitch distributions for: declarations (*top*), questions (*middle*), and exclamations (*bottom*)

We can see that for both speakers the pitch mean/mode and the standard deviation increase for questions and exclamations in comparison to declarations, as is expected. However, while BG's mode increases 24 Hz from questions to exclamations, his average pitch does not, and TA's mean/mode pitch even decreases for exclamations compared to questions. However the height of the mode in exclamations compared to declarations justifies the theorized high H- plateau for exclamations presented in [5].

We can also see that skewness reaches its maximum value in questions, which is due to the typical H% rise in intonation at their ends. These rises are short and steep providing for the long and thin tail towards higher frequencies, which can clearly be seen in the histograms. We can also see that exclamations lack this tails, especially for BG, showing that a H% rise is not typical for exclamations.

IV. Conclusion

A thorough analysis of pitch dynamics was undertaken, the results of which are shown in this paper. The presented results are first of their kind for the Macedonian language and are of significant value for further improvement of the intonation module used in our TTS system.

More specifically, the obtained average values of the standard deviation can be used to give a lower and upper bound for the synthesized pitch contours. Such bounds will differ based on the voice used (male/female), and discourse function of the current utterance, in accordance to the results shown here. Most notably, the mean/mode pitch values for the different phrase groups for the speaker BG, will be directly implemented, because our system uses his voice for the unit inventory. The data gathered for the female speaker TA will also be of value when expanding the system to include a female voice.

References

[1] D. Jurafsky, J. H. Martin, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, *2 ed.*, Prentice Hall, May 2008.

[2] С. Бојковска, Л. Минова-Ѓуркова, Д. Пандев и Ж. Цветковски, „Општа Граматика на Македонскиот Јазик“, Просветно Дело, 2008.

[3] V. Friedman, "Macedonian", *SEELRC*, 2001.

[4] B. Gerazov, G. Shutinoski and G. Arsov, "A Novel Quasi-Diphone Inventory Approach to Text-To-Speech Synthesis", *MELECON '08*, Ajaccio, France, May 5-7, 2008

[5] B. Gerazov, Z. Ivanovski, "Analysis of Intonation in the Macedonian Language for the Purpose of Text-to-Speech Synthesis", *EAA EUROREGIO 2010*, Ljubljana, Slovenia, 15 – 18 Sep, 2010

[6] Rabiner L.R., R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978