Analysis of pitch profiles in Germanic and Slavic languages.

Grazyna Demenko
Adam Mickiewicz University

Bistra Andreeva
Universität des Saarlandes

7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Analysis of pitch profiles in Germanic and Slavic languages

Grażyna Demenko[1]
Bernd Möbius[2]
Bistra Andreeva[2]
[1] Department of Linguistics, Adam Mickiewicz University, Poland
[2] Computational Linguistics & Phonetics, Saarland University, Germany

Summary
This study presents the results of a large-scale analysis of various measures of pitch range and pitch variation in two typologically different language groups: English and German (Germanic) and Bulgarian and Polish (Slavic). The comparison is based on large multi-speaker corpora (48 speakers for Polish, 60 for each of the other three languages). Linear mixed models were computed that include various distributional measures of pitch level, span and variation, revealing characteristic differences across languages and between typological language groups. A classification Multi_Layer Perceptron algorithm based on the relevant parameter measures (span, kurtosis and skewness values for pitch distributions for each speaker) succeeded in separating the typologically different languages with 95% correctness. Significant differences between the language groups were found: German and English speakers use lower pitch maxima, narrower pitch span, and generally less variable pitch than Bulgarian and Polish speakers. Short introduction to multilingual prosody description based on general language aspects is given. Sources of variability of fundamental frequency and their relevance for the extraction and interpretation of paralinguistic information are discussed with relation to multilingual speech prosody processing.

PACS no. xx.xx.Nn, xx.xx.Nn

1. Introduction

Several studies over the past decades have shown that linguistic communities (different social groups within a single language or speakers of different languages) tend to be characterized by particular pitch profiles (pitch range and pitch variation, see Dolson, 1994 for a review). Various cross-linguistic studies also indicate language specific differences with respect to *f0*. Comparing typologically different languages (English, Spanish, Japanese, Tagalog), Hanley et al. (1966) and Hanley and Snidecor (1967) found that the fundamental frequency of English males had the lowest median *f0*. Later studies compared Polish vs. English (Majewski et al., 1972), Mandarin vs. English (Keating and Kuo, 2012), British English vs. German (Mennen et al., 2012), or Russian vs. German (Nebert, 2013). Some studies showed that bilingual speakers differ when speaking their two languages. For example, bilingual English/ Japanese speakers used a higher pitch in Japanese than in English (Graham, 2013). These findings demonstrate that such differences need not be due to physiological differences between speakers of different languages. Ohala and Gilbert's (1979) report on experiments in which listeners can identify their own language (Japanese, Cantonese and English) based solely on prosodic cues (*f0*, amplitude and timing characteristics). It has further been found that some languages are discriminable purely by their fundamental frequency (Ramus and Mehler, 1999 for English and Japanese, Maidment, 1983 for English and French and de Pijper, 1983 for English and Dutch). However, it is difficult to compare the data reported in these publications, because most studies have been limited to either male or female (mostly small numbers of) speakers, the analyses were based on different discourse types, or the methods for *f0* estimation were different.

2. Fundamentals of multilingual prosody description

English and German belong to the West Germanic language family, Bulgarian is a member of the Southern branch of the Slavic language family and Polish belongs to the West Slavic languages. The primary division between the four languages concerns prosodic features: word stress, accent, rhythm and intonation. Word stress is defined as the relative emphasis that may be given to certain syllables in a word by means of greater duration, higher intensity and unreduced spectral properties of the (vocalic) unit. Accent is relevant at a higher level than the individual syllable – namely within a prosodic phrase and is cued primarily by pitch movement (e.g. rising, falling, rising-falling). Intonation refers to the combination of pitch accents and other phrasal level pitch properties such as pitch direction at phrase edges and the relative height of accent peaks.

Different languages are characterized by different speech rhythm. The isochrony of the syllable, the foot and the mora is the basic assumption behind the rhythmic categories. In *stress-timed languages*, the non-stressed syllables are shortened so that the interval between stressed syllables becomes more isochronous than they otherwise would be. This contrasts with languages that have *syllable-timed* or *mora-timed languages*, where each syllable or mora takes roughly the same amount of time regardless of stress. Arthur Lloyd James‘ (1940) metaphorical description of English as sounding like Morse Code and French like a machine gun is given credit for starting the enduring dichotomy between „stress-timed“ and „syllable-timed“ languages.

Bulgarian, English and German are languages with variable (free) stress. The stress assignment in this languages serves as a feature to distinguish otherwise identical words (e.g. *DI*gest vs. di*GEST*), whereas in Polish the stress does not play any distinctive role and the place of stress is nearly always on the penultimate syllable (fixed stress). English and German are said to be stress-timed languages (Abercrombie, 1967, Kohler, 1982), Bulgarian and Polish occupy an intermediate position on a scale of rhythm and are characterized as being of a mixed type (Dauer, 1987, Dimitrova, 1998). All four languages are intonation languages where pitch variation is used for a range of functions such as disambiguation of different syntactic structures, signalling the difference between statements and questions, and between different types of question, indicating the emotional state and attitudes of the speaker, highlighting important elements of the spoken message and regulating conversational interaction. The pitch patterns of speech are systematic and language-specific. Anderson (1979) analyzes the ‘neutral’ pattern in German as a rising pitch on the first accented syllable followed by falling pitch on the second (‘flat hat’). In English, both the first and second accents are rising-falling. The ‘neutral’ pattern in Bulgarian is a low pitch on the first accented syllable followed by falling pitch on the second (Andreeva et al., 2001). In Polish the most frequently realized type of an accent is a static/flat

accent (tones of the accented and post-accented syllable are on almost the same level, difference not exceeding +/-2 semitones). Quite frequent also accented syllable with high pitch (flat accent) followed by falling pitch on post-accented syllable (Demenko, 2014).
In English, German and Polish the Yes/No questions carry a rising tone on the final accented unit. In Bulgarian the question intonation is characterized by a rising-falling nuclear pitch. The pitch maximum of the rise is reached within the vowel of the accented syllable, which is followed by a falling pitch movement. The interval between the highest and the lowest level of the fundamental frequency is an octave at least

3. Basic aspects of fundamental frequency contour processing

3.1. Sources of *f0* parameter variability

Defining and interpretation of pitch patterns requires a separation of sources of variation, which is a very complex task. For example, a specific increase in the fundamental frequency may indicate a grammatical function (e.g., the selection of accent or boundary tone), paralinguistic function (e.g. temporary emotion) or non-linguistic factors. We will shortly discuss four levels of fundamental frequency analysis: 1) non-linguistic (physiological and neuro-physiological), 2) non-linguistic (neurolinguistic), 3) paralinguistic (psycholinguistic and sociolinguistic) and 4) linguistic.

1) Physiological and neurophysiological
Complex functions of the respiratory and articulatory system directly shape the structure of the speech. Range of variability of segmental and suprasegmental parameters, especially fundamental frequency determine: a) physiologically and anatomically factors, b) the external factors environmental (short-term or permanent), and c) the technical and situational circumstances.

a) *Features conditioned physiologically and anatomically*

• Height, weight, gender, individual biomechanical characteristics of the organ of speech.
Characteristics of the speech signal under these conditions are considered rather as a relatively stable, such as the impact of speaker's height, weight, gender at the fundamental frequency confirmed many researchers (Schuller et al., 2013) (Ey et al., 2007).

• Habits of articulation.
There are mainly in relation to the use of a specific range of *f0* parameter changes but also to voicing/unvoicing. Among others, most important factors are not only anatomical and physiological but also psycholinguistic and sociolinguistic.

• Hormonal changes.
Rapid changes in the voice at puberty both boys and girls are evidence of the influence of sex hormones on organ voice and physiologically justified (Hacki, Heitmüller, 1999)

• Aging.
Over the years, natural variations occur in the organ voice. This slow process leads to fundamental changes in the acoustic parameters of voice, for example, the fundamental frequency is reduced in women, in men increase (Ramig, Ringel, 1983)

• Pathologies.
Pathologies are caused by various factors, such as alcohol, drugs, drugs. Regular use of aspirin may cause slight bleeding in the vocal folds, voice deepening and hoarseness. Numerous studies confirm the negative effect of nicotine on the organ of voice, pointing to a permanent change in heavy smokers (Lee et al., 1999). Analysis of a variety of long-term and short-term voice disorders of organic or functional nature based of fundamental frequency variability is used in the early diagnosis of cancer of the larynx (Demenko, 1999).

b) *External physical factors*

• Contamination.
The most common problem is contamination, such as dust, causing allergies and infections. Dust disturbs the respiratory tract, causes swelling and inflammation of the mucous membrane of the nose and throat, which can cause hoarseness or even complete loss of voice

• Humidity.
Too high or too low humidity have a negative impact on the work of the organ voice. In the case of a constant work in such conditions can cause permanent changes in voice (Hemler et al., 1997).

• Noise.
Noise is considered the most important environmental factor affecting the production of speech. To speak in noise, extra effort is needed: speaking louder, higher, with a more careful articulation. Noise can also lead to permanent changes in the speech organ (Biber, 1991; Junqua, 1996).

c) *Situational determinants.*

The physical distance between the speaker and the listener, or a microphone, directly affects the elevation of the speaker voice (Titze, Winholtz, 1993).

2) Neurolinguistic
One of the most important and least studied factors affecting the acoustic structure of speech, is the degree of control (or lack of control) of voice by the speaker. Stress and extreme emotions have a direct impact on the physiology and functioning of the organ voice - changing the settings of articulation and voice track stimulation, which causes specific changes in the structure of the speech signal. The sound-related determinants of stress and personality features can be considered rather constant (Mairesse, Walker, 2006).(Pollermann, 2002)

3) Psycholinguistic and sociolinguistic
Language specific components have also been found to be important in the perception and production of paralinguistic aspects (Loveday, 1981 for politeness in Japanese and English, Chen et al., 2004 for ‘confident’, ‘friendly’, ‘emphatic’ and ‘surprised’ in British English and Dutch). Luchsinger and Arnold (1965) found that Puerto Rican girls in New York City and native American women use *f0* differently. While Puerto Rican girls tend to speak on a rather high pitch, many American women prefer to speak on a low pitch level. Dialects of a language can also differ with respect to the use of *f0* (e.g. Deutsch et al, 2009, Torgerson, 2011). Different aspects of socio-cultural, socio-economic speaker’s status have high significance in forensic speaker recognition (Nolan, 2007)

4) Linguistic level
Analysis of the fundamental frequency differences is particularly useful for the characterization and recognition of speaker’s language competence. Pitch patterns provide basic information on grammatical and syntactical language dependant relations. (Hirst, Di Cristo, 1998).

3.2. Methodological and technical problems

Considering the practical use of fundamental frequency processing in speech technology must account the basic determinants of sources of variability: (a) individual (features-individual speaker at various levels), (b) the external environment, forced (e.g., Lombard, cocktail party, spatial factors, atmospheric) and c) technical (microphone, acoustic track, environmental conditions).

Methodological and technical problems *f0* parameter processing will include: 1) analysis of temporal variability, 2) reliable extraction, 3) normalization, and 4) correlations between different sources of variability.

1) Temporal variations

In general characteristic features of speech features most often referred to as short-, medium-and long-term (Demenko, 1999).(Schuller et al., 2013)

• Long-term. Biological factors related to the size, weight, age, gender speakers considered to be relatively stable. In addition, language habits resulting from membership in social, education, dialect, personality, habits articulation of specific, idiosyncratic pitch patterns structures help to define relatively effective characteristics of the speaker (Walton, Orlikoff, 1994).

• The medium-term. Features caused by more or less temporary health conditions, such as drowsiness (Krajewski et al., 2009), poisoning, e.g., an alcohol (Pisoni, Martin, 1989); (Schiel, 2012) a general medical condition (Maier et al., 2009) mood (Ellgring, Scherer, 1996) are usually temporary.

• Short. Most often associated with temporary changes to the way of expression, caused e.g. emotions (Schuller et al., 2009), stress (Hansen, 1996), uncertainty (Litman et al., 2009), politeness (Yildirim et al., 2005), frustration (Lee, 2001), sarcasm (Rankin et al., 2009), physical ailments (Cowin et al., 2003), are unstable, however, cause significant, temporary, specific changes in segmental but mostly in suprasegmental (fundamental frequency) patterns.

2) Reliable extraction

Reliable extraction of fundamental frequency of speech in itself is a challenge. For example, in the telephone speech signal parameter *f0* is often out of band transmit narrowband networks (0.3 to 3.4 kHz) and algorithms must be based exclusively on data from higher harmonics (Hess, 1982).

The complexity of the problem is caused by the specific characteristics of the speech signal:
(1) The excitation signal is quasi-periodic. It has irregular pitch changes, even for average quality voices there are significant abnormal signal periodicity.
(2) Voice track shape changes at intervals of the order of ms, which causes significant variation in the structure of the spectrally-time signal.
(3) The signal is discontinuous, existing pauses in

speech and unvoiced consonants are the cause of interruptions in the course of the fundamental frequency.
(4) Speech fundamental frequency can be varied within the range of about 4 octaves (50 - 800 Hz).
(5) The speech signal can be further distorted (clipped lower frequency range, high noise level). Different techniques have their specific drawbacks of the method used signal processing (e.g., sensitivity to the changes in signal level).
Errors usually includes extraction into one of three groups:
a) the so-called large errors (often due to improper measurement of second or higher harmonics
b) minor errors that result from the inaccuracy of the method used,
c) errors in the detection of voicing/unvoicing.

The optimal solution currently used is the use of several parallel working extractors fundamental frequency and taking into account statistically the most reliable combination of results.
General variability of the fundamental frequency is controlled by the speaker (e.g., resolution, type accents), while microfluctuations signal (arising out of phonetic context) are determined by aerodynamic phenomena.
In fact, the human auditory system smoothes out the irregularities and changes the parameter *f0* perceives as a continuous melodic structure. Microprosodic fundamental changes of *f0* parameter and pauses caused unvoicing consonants have no effect on auditory perception accent, while contributing to the impression of naturalness signal.

3.3. Normalization

Fundamental frequency curves may exhibit differences in terms of:

a) the continuity/discontinuity
b) rate of speech and rhythm features
c) different distribution of extremes (specific location of pitch accents).

One way to achieve invariance is to normalize the data. The normalization and standardization is an important aspect of the preparation of data for analysis, as already simple scaling the coordinates may lead to a different division into groups. The most common methods include normalization parameter change *f0* with respect to arbitrarily chosen values.
Commonly used normalization using the mean parameters of the distributions of the fundamental frequency, the mean and standard deviation (Rose, 2002). One of the reasons for the difficulties in modeling and analysis of the structures of intonation is uneven speech rate of speech. Numerous attempts to solve this problem to the problem of normalization are based on nonlinear method (time warping), dynamic programming technique (DTW).

3.4. Correlating features

Among the features there are significant correlations. For example, certain demographic, dialectological ethnicity are significantly related to each other, determine the accent, the pronunciation, so-called. sociolect, which proved to be extremely useful for example in forensic (Becker et al., 2008).

4. Experimental data and measurement

4.1. Data

Two Slavic (Bulgarian and Polish) and two Germanic (German and British English) languages are in the focus of this study (cf. 2.). The material analyzed is continuous read speech taken from two comparable multi-lingual speech databases, for German and English: EUROM-1 (Chan et al., 1995) and for Bulgarian and Polish: BABEL (Roach et al., 1998). We used a subset of the data, consisting of 3 cognitively linked short passages, containing 5 thematically connected sentences, read by 60 speakers (30 male and 30 female) for Bulgarian, German and English and 48 speakers (24 male and 24 female) for Polish. The passages were based on identical, real-life topics for the different languages, freely translated and adapted for Bulgarian, German and Polish from the original English texts. The overall length of the analyzed material is about 70 minutes for Polish and 90 minutes for each of the other three languages.

4.2. *f0* Measures

Pitch values were collected at 0.01 seconds time steps for the male and 0.005 seconds time steps for the female speakers using the RAPT algorithm (Talkin, 1995) implemented in the program ‘get_f0’ from the ESPS software package. The automatically extracted *f0* values were verified and manually corrected, if necessary. Irregular voiced stretches of speech due to laryngealization were excluded from further analyses.

According to Ladd (1996), *f0* values can be attributed to two partially related but distinct characteristics of a speaker's performance: (a) pitch level, i.e. the overall height of the speaker’s voice, and (b) pitch span, i.e. the range of frequencies covered by the speaker. To analyze the cross-language differences in pitch range and variation, the following distributional measures were calculated: mean *f0* values for level and the pitch excursion for span, whereas the latter was simply computed as the difference between maximum and minimum pitch values over a passage. The obtained Hertz measurements for span were additionally converted to semitones by means of the formula (Reetz, 1999):

39.863 * log10(Maximum/Minimum).

The measures describing the variation and shape of the *f0* distribution were standard deviation (SD), kurtosis and skewness (in Hz).

Means and standard deviations for the distributional measures used for language comparison of level and span, by language and gender are given in Table I.

5. Statistical evaluations

As first evaluation of representativeness of speech data preliminary analysis of stability of *f0* distributions for each speaker has been analysed. Each distribution was based on three passages approx. 25-35 seconds long (5 sentences). The average difference between mean *f0* values of three distributions was approx. 5 Hz. Figure 1 shows three distributions for one male English speaker.

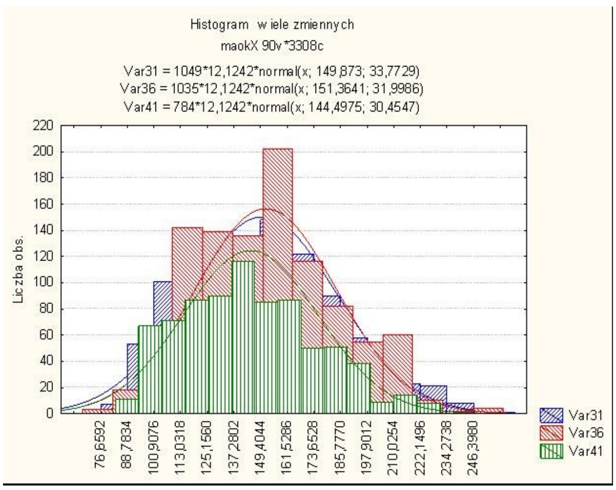

Figure 1. *f0* distributions from 3 passages for one speaker.

In order to determine the influence of the speakers‘age on the *f0* values under investigation (cf. 3.1.) first the speakers were divided into two age groups: (a) younger than 32 years and (b) older than 32 years. Subsequently, linear mixed models with the respective *f0* measure as dependent factor, speaker and item as random factors and language (Bulgarian/Polish/English/German), gender (male/female), age (younger speakers/older speakers) and height as independent factors, as well as all their possible interactions, were computed for each dependent variable in separate analyses (for details see Andreeva et al., 2014).

Separate Tukey post-hoc tests were carried out per variable, if appropriate. The confidence level was set at α=0.05.

5.1. The effect of age and height

The mean, median, standard deviation and range values of the age and height of the female and male speakers are reported in Tables IIa, IIb, IIIa and IIIb.

Table IIa. Mean median, standard deviation (SD) and range values of the age of female speakers.

	BG	PL	DE	EN
median	36.0	25.5	27.0	32.5
mean	36.0	32.9	30.4	36.1
SD	13,0	13.4	8.6	11.3
max	69	57	61	56
min	20	19	22	19

Table IIb. Mean median, standard deviation (SD) and range values of the age of male speakers.

	BG	PL	DE	EN
median	36.0	24.0	26.0	29.5
mean	35.2	29.8	30.0	34.7
SD	12,3	12.5	7.4	12.4
max	61	60	54	66
min	21	21	23	21

Table IIIa. Mean median, standard deviation (SD) and range values of the height of female speakers.

	BG	PL	DE	EN
median	164.5	165.0	170.0	162.0
mean	164.1	166.2	166.2	164.3
SD	6,2	7.1	4.4	7.6
max	176	180	183	182
min	145	152	160	153

Table IIIb. Mean median, standard deviation (SD) and range values of the height of male speakers.

	BG	PL	DE	EN
median	180.0	182.0	182.0	178.5
mean	180.5	181.5	181.4	179.8
SD	5,6	7.1	7.0	6.4
max	194	196	195	200
min	170	170	162	170

The results of the statistical analysis show a significant main effect for age on minimum *f0*, span in Hertz and semitones and SD. The 'older' speakers had a significantly lower minimum *f0* (F [1, 195.9] = 17.39, p<0.001), higher *f0* span in Hertz (F [1, 196] = 8.39, p<0.05) and semitones (F [1, 196] = 20.90, p<0.001) and higher SD (F [1, 196] = 4.85, p<0.05) than the 'younger' speakers (cf. Figure 2).
There was no significant main effect for speakers' height.

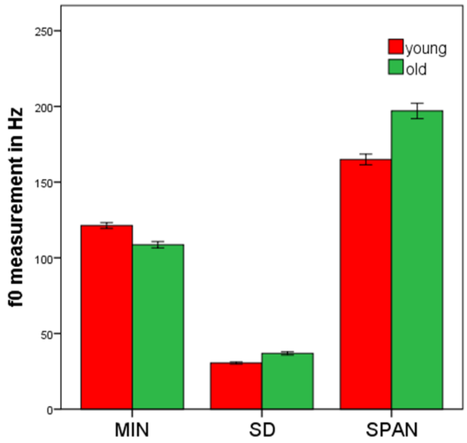

Figure 2: Age main effect on *f0* minimum, SD and span in Hertz.

5.2. The effect of gender

Predictably, gender had a significant main effect on mean *f0* (F [1, 220] = 1143.382, p<0.001), minimum *f0* (F [1, 220] = 669.243, p<0.001), maximum *f0* (F [1, 220] = 807.7228, p<0.001), *f0* span measured in Hz (F [1, 220] = 270.4249, p<0.001), SD (F [1, 220] = 202.9187, p<0.001) and skewness (F [1, 220] = 7.8404, p<0.0056), with females having significantly higher *f0*. Gender did not differ in kurtosis and *f0* span measured in semitones (see Figure 3 for *f0* mean, maximum, minimum and span measured in semitones).

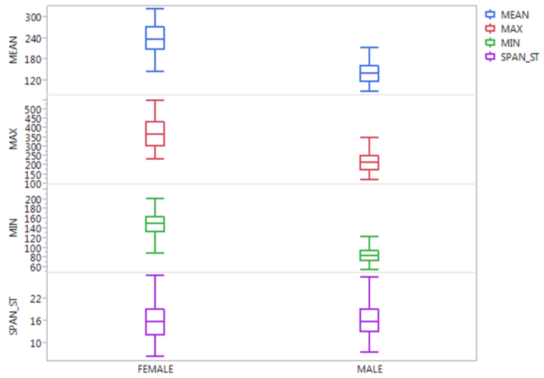

Figure 3: Gender main effect on *f0* mean, maximum, minimum and span in semitones

5.3. The effect of language

However, over and above the expected gender effect, there was also a significant main effect of language on all measurements except on minimum *f0,* where the speakers are near the floor of their physiological *f0* range. Separate post-hoc tests showed that Bulgarian and Polish speakers had a significantly higher mean *f0* (F [3, 220] = 87.9677, p<0.001) and *f0* span in semitones (F [3, 220] = 41.1905, p<0.001) than English and German speakers. In the Slavic languages *f0* varies most strongly (possibly indicating more liveliness). Polish and Bulgarian reveal significantly higher SD values than English and German, although the English values are significantly greater than the German ones (F [3, 220] = 60.7884,

Table I. Means and standard deviations for the distributional measures, by language and gender. The values for each measure are given in Hz except for the second span measure which is in semitones.

Measure	Bulgarian		Polish		German		English	
	male	female	male	female	male	female	male	female
mean	160 (21)	272 (32)	163 (22)	266 (24)	118 (16)	210 (20)	128 (22)	217 (20)
minimum	88 (15)	149 (25)	85 (15)	149 (21)	80.0 (12)	146 (25)	84 (13)	151 (23)
maximum	238 (37)	422 (52)	260 (37)	443 (62)	176 (29)	299 (31)	200 (43)	337 (53)
span	150 (37)	273 (49)	176 (36)	294 (66)	96 (26)	154 (35)	116 (39)	186 (61)
span (s.t.)	17.2(3.6)	18.2 (3.0)	19.5 (3.4)	18.9(3.7)	13.6(2.7)	12.7 (3.5)	14.9 (3.4)	13.9 (4.1)
SD	29 (8)	52 (12)	32 (8)	53 (14)	17 (5)	28 (7)	22 (9)	35 (11)
skewness	.01 (.33)	.17 (.33)	.11 (.53)	.47 (.43)	.41 (.45)	.29 (.31)	.54 (.42)	.66 (.46)
kurtosis	-.26 (.48)	-.19 (0.46)	.29 (.54)	.28 (.86)	.30 (.93)	-.17 (.75)	.34 (1.00)	.51 (1.15)

p<0.001). The four languages differ significantly in their maximum *f0* values (F [3, 220] = 90.5398, p<0.001). We found a positively skewed *f0* distribution for the four languages. This implies that the most frequent *f0* observation occurs lower than the mean. The skewness values for English speakers were significantly higher than those for German and Polish speakers and the values for the German and Polish speakers were significantly higher than those for Bulgarian speakers (F [3, 220] = 21.3182, p<0.001). English speakers had a higher kurtosis than German and Bulgarian speakers, and Polish speakers had a higher kurtosis than Bulgarian speakers (F [3, 220] = 13.1106, p<0.001). This reflects the fact that *f0* in Bulgarian and German is distributed over a narrower area (cf. Table I and Figure 4).

The statistical analysis further revealed a significant interaction between language and gender for mean *f0*, maximum *f0*, SD, and skewness. This interaction can be explained by the higher *f0* register used by the Slavic speakers compared to the German speakers. The (relatively high) register for Polish and Bulgarian male speakers is in the same range of absolute *f0* values as that of English and German female speakers, causing them to group together in some analyses. Thus, the general pattern of higher *f0* values for the Slavic speakers than for Germanic speakers is retained. These results are in line with our findings in Andreeva et al., 2014. Table IV shows the *f0* measure patterns by languages.

Table IV. *Language-group differences for the f0 measures on the basis of Tukey post-hoc comparisons.*

***f0* measure**	**significant language-group differences**
mean *f0*	BG = PL > EN = DE
min *f0*	N.S.
max *f0*	PL > BG > EN > DE
span s.t.	PL = BG > EN = DE
SD	PL = BG > EN > DE
skewness	EN > DE = PL > BG
kurtosis	EN = PL > PL = DE > BG

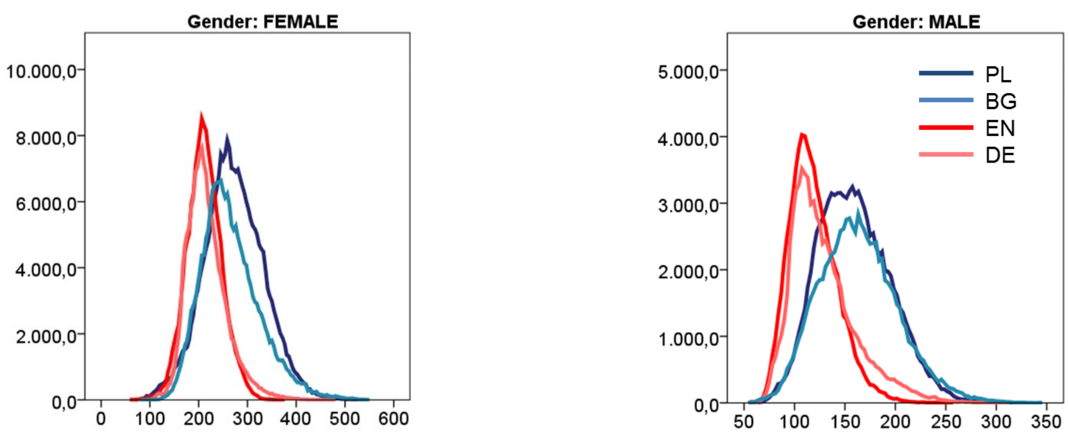

Figure 4: Probability density function for female (left panel) and male (right panel) speakers

6. Classification with Multi-Layer Perceptrons (MLPs)

As discussed above, most *f0* measures - with the exception of minimum *f0*, which tends to reflect the lower physiological limit of *f0* production and is therefore quite stable across languages - appear to be characteristic of individual languages. However, the more general pattern that emerges from Table IV is a separation of languages along the line of typologically distinction, in that the Slavic languages (Bulgarian and Polish) as a group differ from the Germanic languages (English and German) for most *f0* measures in a consistent manner.

To estimate the strength of the contribution of *f0* measures to the typological distinction and the possibility of distinguishing data which were not linearly separable by classical statistical methods, a classification with MLP (with backpropagation learning algorithm) was performed. The MLP net model was used as an attempt to explain the categorical variable (a) “language group” (Slavic vs. Germanic) and (b) “gender” (male vs. female).

A Multi-Layer Perceptron with 3 input neurons equalling the number of input features (span, kurtosis and skewness), 7 hidden layers, and 2 output neurons for each language group was used because a performance maximum was observed using 7 neurons in the hidden layers compared to other nets

architectures (20 different nets were used for preliminary evaluation of the quality of training). The outputs were normalized as posteriors by a softmax function. For the training 70% of the data were used, the validation set and the test set comprises 15% of the data.
The classification was based on three variables, *f0* span (in semitones), *f0* kurtosis and skewness. These three variables were selected because they are representative of pitch range and pitch variability, respectively, and there were no interactions found between gender and language in the statistical analysis with the linear mixed models. Since the classification was carried out for male and female speakers together we expect these two variables to be key ingredients of language (group) specific pitch profiles.
The graph in Figures 5 provides a visual representation of *f0* span (in semitones) and kurtosis (in Hz) for the Slavic and Germanic language groups. The figure show a clear separation between the different groups – 91 % correct classification for the Germanic and 81 % for the Slavic language group.

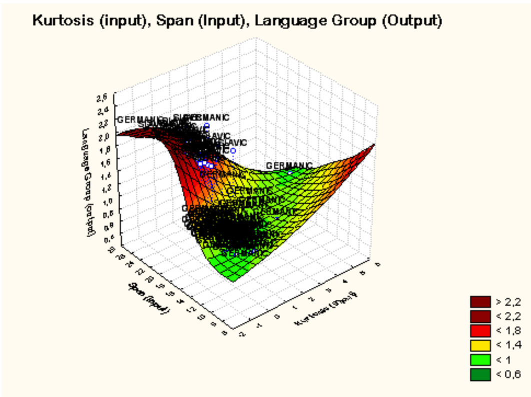

Figure 5: Visualization of language group classifycation (Slavic vs. Germanic) on the basis of *f0* span and kurtosis.

The English and German speakers cluster in the lower right corner of the span/kurtosis plane, while the Bulgarian and Polish speakers cluster mostly in the higher left sector. Further research using different methods and measures of analysis is needed to explain this pattern. The measures used in this study are to general for the precise interpretation of the results. An alternative to measuring *f0* distribution is to reduce the *f0* contour to a series of target points representing the significant pitch changes by automatic stylization (cf. Campione and Véronis, 1998) or by pitch accents labelling. The classification male/female was also based on three variables, *f0* span (in semitones), *f0* kurtosis and skewness.

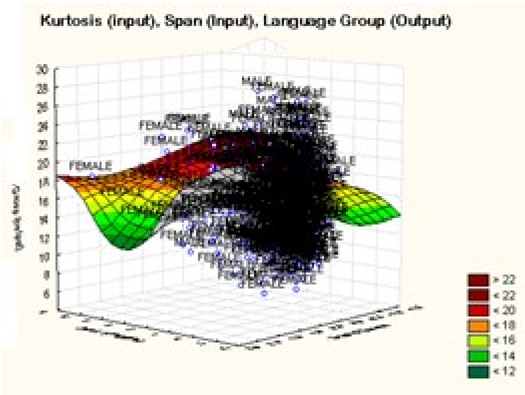

Figure 6: Visualization of gender classification (males vs. females) on the basis of *f0* span and kurtosis.

The graph in Figure 6 provides a visual representation of *f0* span (in semitones) and kurtosis (in Hz) for the males and females. The figure shows a no separation between the gender – 48 % correct classification for males and 35% for females**.**

7. Discussion and Conclusions

This paper contributes to the growing number of studies on cross-language differences in pitch range and pitch variation. Our results are in line with our previous research (Andreeva et al. 2014) and confirm the hypothesis that linguistic communities tend to be characterized by particular pitch profiles. The male and female speakers of the Slavic group used considerably higher mean, maximum *f0* and span in semitones and showed a larger SD (possibly indicating more liveliness) than the speakers in the Germanic group. Classification with Multi-Layer Perceptron with span, kurtosis and skewness as input variables show clear separation between the Germanic and Slavic group.
In future work we expect to refine our measures of pitch range, by including linguistically based measures which were found to be better predictors of differences in pitch range and pitch variation across speakers and languages (Campione and Véronis, 1998, Mennen et al., 2012), and also by adding data from Bulgarian and Polish L2 speakers of English and German, more languages, as well as spontaneous speech data.

Acknowledgements

We would like to thank Ryszard Gubrynowicz (Speech Acoustics Laboratory, Institute of Fundamental Technology Research, Polish Academy of Science) and Snezhina Dimitrova (English Department, Sofia University “St. Kliment Ohridski”) for kindly providing the Babel databases for Polish and Bulgarian, respectively.

References

[1] Abercrombie, D.(1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

[2] Anderson, K. O. (1979). On the contrastive phonetics of English and German intonation. *Festschrift für Otto von Essen anlässlich seines 80. Geburtstages*. In H-H. Wängler (ed.), Hamburger Phonetische Beiträge 25, pp. 25-35.

[3] Andreeva, B., Avgustinova, T and Barry, W.J. (2001). Link-associated and focus-associated accent patterns in bulgarian. In: Gerhild Zybatow, Uwe Junghanns, Grit Mehlhorn and Luka Szucsich (eds.), *Current Issues in Formal Slavic Linguistics*, volume 5 of Linguistik International, Peter Lang: Frankfurt am Main, pp. 353-364.

[4] Andreeva, B., Demenko, G., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., Jastrzebska, M., and Trouvain, J. (2014). Comparison of pitch range and pitch variation in Slavic and Germanic languages. In Proc. Speech Prosody 2014, Dublin.

[5] Becker, T., Jessen, M., Grigoras, C. (2008). *Forensic speaker verification using formant features and Gaussian mixture models.* Proceedings of the Interspeech, pp. 1505-1508.

[6] Biber, D. (1991). *Variation across speech and writing*: Cambridge University Press.

[7] Campione, E., and Véronis, J. (1998). A statistical study of pitch target points in five languages. *Proceedings of ICSLP’98*, pp. 1391-1394.

[8] Cowin, L., Davies, R., Estall, G., Berlin, T., Fitzgerald, M., Hoot, S. (2003). De-escalating aggression and violence in the mental health setting. *International Journal of Mental Health Nursing, 12*(1), pp. 64-73.

[9] Chan, D. Fourcin, A.; Gibbon, D.; Granstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouropoulos, J.; Senia, F.; Trancoso, I.; Veld, C. and Zeiliger, J. (1995). Eurom - a spoken language resource for the EU. In *Eurospeech’ 95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology*, 1, Madrid., 18-21 September 1995, pp. 867–870.

[10] Chen, A, Gussenhoven, C., and Rietveld, T. (2004). Language-specificity in the perception of paralinguistic intonational meaning, *Language & Speech* 47, pp. 311–349.

[11] Cowin, L., Davies, R., Estall, G., Berlin, T., Fitzgerald, M., Hoot, S. (2003). De-escalating aggression and violence in the mental health setting. *International Journal of Mental Health Nursing, 12*(1), pp. 64-73.

[12] Dauer, R. (1987). *Phonetic and phonological components of language rhythm*, Proceedings of the XIth International Congress of Phonetic Sciences, Talinn, Estonia, pp. 447-450.

[13] Dimitrova, S. (1998). Bulgarian Speech Rhythm: Stress- Timed or Syllable-Timed?, *Journal of the International Phonetic Association*, vol. 27.1, pp. 27-33.

[14] Demenko, G. (1999). Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy. Poznań: Wydawnictwo Naukowe UAM.

[15] Demenko, G. (2014), Corpus speech processing, ed.Exit,Warszawa, in print

[16] de Pijper, J. R. (1983). *Modelling British English intonation*, Dordrecht - Holland: Foris.

[17] Deutsch, D., Le, J., Shen, J., and Henthorn, T. (2009). The pitch levels of female speech in two Chinese villages. *Journal of the Acoustical Society of America*, April, 125, EL208.

[18] Dolson, M. (1994). The pitch of speech as a function of linguistic community, *Music Perception* 11 (3), pp. 321–331.

[19] Ellgring, H., Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior, 20*(2), pp. 83-110.

[20] Ey, E., Pfefferle, D., Fischer, J. (2007). Do age-and sex-related variations reliably reflect body size in non-human primate vocalizations? A review. *Primates, 48*(4), pp. 253-267.

[21] Graham, C., (2013). Revisiting f0 Range Production in Japanese-English Simultaneous Bilinguals. *Annual Report of UC Berkeley Phonology Lab*, 110–125.

[22] Hacki, T., Heitmüller, S. (1999). Development of the child’s voice: premutation, mutation. *International journal of pediatric otorhinolaryngology, 49*, pp. S141-S144.

[23] Hansen, J. H. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech communication, 20*(1), pp. 151-173.

[24] Hanley, .D. and Snidecor, J.C. (1967). Some acoustic similarities among languages, *Phonetica* 17, pp. 141–148.

[25] Hemler, R. J., Wieneke, G. H., Dejonckere, P. H. (1997). The effect of relative humidity of inhaled air onacoustic parameters of voice in normal subjects. *Journal of Voice, 11*(3), pp. 295-300.

[26] Hess, W. J. (1982). Algorithms and devices for pitch determination of speech signals. *Phonetica, Hess, W. J. (1982). Algorithms and devices for pitch determination of speech signals.* (39(4-5)), pp. 219-240.

[27] Hirst, D., Di Cristo, A. (1998). *Intonation Systems: A Survey of Twenty Languages*: Cambridge University Press.

[28] Junqua, J.-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication, 20*(1), pp. 13-22.

[29] Keating, P. & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and

Mandarin, *Journal of the Acoustical Society of America* 132, pp. 1050–1060.

[30] Kohler, K. (1982) *Rhythmus im Deutschen*, Arbeitsberichte, Institut für Phonetik der Universität Kiel, 19, pp. 89-106.

[31] Krajewski, J., Batliner, A., Golz, M. (2009). Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods, 41*(3), pp. 795-804.

[32] Ladd, D.R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

[33] Lee, D. (2001). *Cognitive linguistics: An introduction* (Vol. 13): Oxford University Press Australia.

[34] Lee, L., Stemple, J. C., Geiger, D., Goldwasser, R. (1999). Effects of environmental tobacco smoke on objective measures of voice production. *The Laryngoscope, 109*(9), pp. 1531-1534.

[35] Litman, D. J., Rotaru, M., Nicholas, G. (2009). *Classifying turn-level uncertainty using word-level prosody*. Proceedings of the Interspeech, pp. 2003-2006.

[36] Lloyd James, Arthur. 1940. *Speech signals in telephony*. London: Pitman & Sons.

[37] Loveday, L. (1981). Pitch, politeness and sexual role: an explanatory investigation into the pitch correlates of English and Japanese formula,. *Language and Speech* 24, pp. 71–89.

[38] Luchsinger, R. and Arnold, G. (1965). *Voice-Speech-Language*. Constable&Co Ltd., London.

[39] Maidment, J. A. (1983). Language recognition and prosody: further evidence, *Speech, hearing and language: Work in progress*, University College London 1, pp. 133–141.

[40] Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E. (2009). PEAKS–A system for the automatic evaluation of voice and speech disorders. *Speech Communication, 51*(5), pp. 425-437.

[41] Mairesse, F., Walker, M. (2006). *Automatic recognition of personality in conversation.* Proceedings of the Human Language Technology Conference of the NAACL, pp. 85-88.

[42] Majewski, W., Hollien, H., and Zalewski, J. (1972). Speaking fundamental frequency of Polish adult males, *Phonetica* 25, pp. 119–125.

[43] Mennen, I., Schaeffler, F., & Docherty, G. (2012). Cross-language differences in fundamental frequency range: a comparison of English and German *Journal of the Acoustical Society of America* 131(3), pp. 2249–2260.

[44] Nebert, Augustin Ulrich (2013). *Der Tonhöhenumfang der deutschen und russischen Sprechstimme. Vergleichende Untersuchung zur Sprechstimmlage.* Hallesche Schriften zur Sprechwissenschaft und Phonetik, Band 46. Frankfurt/M.

[45] Nolan, F. (2007). Intonation in speaker identification: an experiment on pitch alignment features. *International Journal of Speech Language and the Law, 9*(1), pp. 1-21.

[46] Ohala, J. J., and Gilbert, J. B. (1979). Listeners' ability to identify languages by their prosody, in P. Léon and M. Rossi (eds.) *Problèmes de Prosodie*, Didier, Ottawa, pp. 123–131.

[47] Pisoni, D. B., Martin, C. S. (1989). Effects of Alcohol on the Acoustic-Phonetic Properties of Speech: Perceptual and Acoustic Analyses. *Alcoholism: Clinical and Experimental Research, 13*(4), pp. 577-587.

[48] Pollermann, B. Z. (2002). *A place for prosody in a unified model of cognition and emotion.* Proceedings of the International Conference on Speech Prosody 2002, pp. 17-22.

[49] Ramig, L. A., Ringel, R. L. (1983). Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech, Language and Hearing Research, 26*(1), pp. 22.

[50] Ramus, F., and Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America* 105 (1), pp. 512–521.

[51] Rankin, K. P., Salazar, A., Gorno-Tempini, M. L., Sollberger, M., Wilson, S. M., Pavlic, D., Miller, B. L. (2009). Detecting sarcasm from paralinguistic cues: anatomic and cognitive correlates in neurodegenerative disease. *Neuroimage, 47*(4), pp. 2005-2015.

[52] Reetz, H. (1999): *Artikulatorische und akustische Phonetik*. Wissenschaftlicher Verlag, Trier.

[53] Roach, P., Arnfield, S., Barry, W.J., Dimitrova, S., Boldea, M., Fourcin, A., Gonet, W., Gubrynowicz, R., Hallum, E., Lamel, L., Marasek, K., Marchal, A., Meister, E., Vicsi, K. (1998). Babel: a database of Central and Eastern European languages, *Proceedings of the First International Conference on Language Resources and Evaluation*, vol. 1, 28-30 May 1998, Granada, Spain, pp. 371–374.

[54] Rose, P. (2002). Forensic speaker identification - Introduction: CRC Press.

[55] Schiel, F., Heinrich, C.,Barfüsser, S (2012). Alcohol language corpus: the first public corpus of alcoholized German speech. *Language resources and evaluation, , 46(3)*, pp. 503-521.

[56] Schuller, B., Steidl, S., Batliner, A. (2009). *The INTERSPEECH 2009 emotion challenge.* Proceedings of the Interspeech, pp. 312-315.

[57] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S. (2013). Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer Speech & Language, 27*(1), pp. 4-39. doi: http://dx.doi.org/10.1016/j.csl.2012.02.005

[58] Talkin, D. (1995). A Robust Algorithm for Pitch Tracking (RAPT).Kleijn, W. B. and Paliwal, K. K. (eds.), *Speech Coding and Synthesis*. New York: Elsevier.

[59] Torgerson, R. C. (2005*). A comparison of Beijing and Taiwan Mandarin tone register: An acoustic analysis of three native speech styles*, master's thesis, Brigham Young University,

[60] Titze, I. R., Winholtz, W. S. (1993). Effect of microphone type and placement on voice

perturbation measurements. *Journal of Speech, Language and Hearing Research, 36*(6), pp. 1177.

[61] Walton, J. H., Orlikoff, R. F. (1994). Speaker race identification from acoustic cues in the vocal signal. *Journal of Speech, Language and Hearing Research, 37*(4), pp. 738.

[62] Yildirim, S., Lee, C. M., Lee, S., Potamianos, A., Narayanan, S. (2005). *Detecting Politeness and frustration state of a child in a conversational computer game.* Proceedings of the Interspeech, pp. 2209-2212.