Cross-language differences in fundamental frequency range: A comparison of English and German[a)]

Ineke Mennen[b)]
ESRC Centre for Research on Bilingualism and School of Linguistics and English Language, Bangor University, 37-41 College Road, Bangor, LL57 2DG, United Kingdom

Felix Schaeffler
Clinical Audiology, Speech and Language Research Centre (CASL), Queen Margaret University, Edinburgh, EH 21 6UU, United Kingdom

Gerard Docherty
School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

(Received 21 June 2011; revised 28 December 2011; accepted 31 December 2011)

This paper presents a systematic comparison of various measures of f0 range in female speakers of English and German. F0 range was analyzed along two dimensions, level (i.e., overall f0 height) and span (extent of f0 modulation within a given speech sample). These were examined using two types of measures, one based on "long-term distributional" (LTD) methods, and the other based on specific landmarks in speech that are linguistic in nature ("linguistic" measures). The various methods were used to identify whether and on what basis or bases speakers of these two languages differ in f0 range. Findings yielded significant cross-language differences in both dimensions of f0 range, but effect sizes were found to be larger for span than for level, and for linguistic than for LTD measures. The linguistic measures also uncovered some differences between the two languages in how f0 range varies through an intonation contour. This helps shed light on the relation between intonational structure and f0 range. © *2012 Acoustical Society of America*. [DOI: 10.1121/1.3681950]

PACS number(s): 43.70.Kv, 43.70.Fq, 43.70.Jt, 43.70.Bk [AL]   Pages: 2249–2260

I. INTRODUCTION

A key issue in the study of speech prosody is to understand the dimensions along which languages differ and the factors which are responsible for such differentiation. The aim of this paper is to shed light on one aspect of cross-language differentiation, pitch range, which has largely escaped detailed attention to date.

Pitch range is regularly invoked by investigators operating in two domains; first, in the development and testing of models or theories of intonation, and second, in diverse studies investigating factors extraneous to intonation which nevertheless affect the phonetic realization of f0. An example of the former would be studies looking at the scaling of tones in the phonetic realization of intonation, where a central issue is how to capture tonal invariance in the face of variation in pitch range (e.g., Dilley and Brown, 2007; Hirschberg and Ward, 1992; Liberman and Pierrehumbert, 1984). Examples of the latter include clinical studies investigating the extent to which various clinical populations have atypical prosody (e.g., Diehl, Watson, Bennetto, Mcdonough, and Gunlogson, 2009; Hubbard and Trauner, 2007), studies of the vocal correlates of affect (e.g., Banse and Scherer, 1996; Ladd, Silverman, Tolkmitt, Bergmann, and Scherer, 1985; Sobin and Alpert, 1999), and studies of various speaker-oriented factors such as the effects of age, gender, height and weight, ethnicity, and regional accent on f0 (Chen, 2005; Deutsch, Le, Shen, and Henthorn, 2009; Hollien, Hollien, and de Jong, 1997; Nishio and Niimi, 2008; Van Bezooijen, 1995; Van Dommelen and Moxness, 1995).

But in studies such as these, the definition of pitch range is usually not examined critically or presented as anything other than uncontroversial. However, investigators employ a range of measures of pitch range (including, for example, f0 standard deviation, difference between maximum and minimum f0, 90% range, 80% range, quantile measures—see Patterson, 2000,[1] for an overview), suggesting that this may not be as straightforward a concept as might first appear to be the case. It is also the case that in the majority of these studies, pitch range is used as a term for what is probably best referred to as "f0 range," as the studies themselves are not focused on the perceptual correlates of a particular f0 distribution. Henceforth the term "f0 range" is used to denote measures of range within speech performance. Note that by this we are not referring to a speaker's vocal range (i.e., the range of fundamental frequencies which it is physically possible for a speaker to produce), rather to the f0 range deployed in spoken

[a)]Portions of this work were presented in "Pitching it differently: A comparison of the pitch ranges of German and English speakers," Proceedings of the 16th International Congress of Phonetic Sciences, Saarbruecken, Germany, August 1997, and "A methodological study into the linguistic dimensions of pitch range differences between German and English," Proceedings of the Fourth Conference on Speech Prosody, Campinas, Brazil, May 2008.

[b)]Author to whom correspondence should be addressed. Electronic mail: i.mennen@bangor.ac.uk communication, often referred to by many investigators as speaking fundamental frequency or SFF (Baken and Orlikoff, 2000, p. 168). A further problem with many studies on f0 range is that they treat it as a unitary measure (although there are some early attempts to go beyond this, e.g., Eady, 1982) and often do not explicitly distinguish between two aspects of f0 range which by some researchers (see for a discussion Ladd, 2008) are seen as constituting different (and quasi-independent) aspects of a speaker's range; namely f0 level (or "register," the relative height of habitual f0), and f0 span (the extent of spatial differentiation of the high and low ends of a speaker's f0 realizations).

Patterson (2000) is one of very few studies to have treated f0 range as the central object of study. With the aim of developing a unified model of f0 range variation capable of being applied across a number of domains, Patterson evaluated a variety of different f0 range measures. These were broadly classified as two types; long-term distributional (LTD) measures based on an analysis of the f0 distribution within a speaker's performance, and "linguistic" measures, where measures of span and level are linked to specific landmarks in the f0 contour (such as accentual peaks, post-accentual valleys, final low, etc.) which, in turn, are thought to be linked to phonological tones and therefore linguistic in nature (e.g., Ladd, 2008; Liberman and Pierrehumbert, 1984; Pierrehumbert, 1980). Patterson (2000) showed that linguistic measures better characterise perceived f0 range than the more commonly used LTD measures. Specifically, they were shown to be more perceptually valid in that they correlated better with listener judgments of speaker characteristics.

If these measures are capable of characterizing cross-speaker differences, and if some of them, at least, draw on language-specific intonational landmarks, then they may be well-placed for helping to understand the basis of cross-language differences in f0 range which have occasionally surfaced in the literature (e.g., Altenberg and Ferrand, 2006; Dolson, 1994; Hanley, Snidecor, and Ringel, 1967; Keating and Kuo, 2010; Majewski, Hollien, and Zalewski, 1972). In the absence of organic factors which could potentially be the source of such differences (e.g., body size or race-based vocal tract differences, Awan and Mueller, 1996), investigators have attributed cross-language differences to either linguistic or cultural factors, but there has been little attempt to characterize the phonetic basis of these differences (Deutsch *et al*., 2009) nor to question the suitability of the f0 measures used. For example, Majewski *et al*. (1972) found higher mean f0 in young Polish males compared to previously reported values for American males. As they were unable to find a significant relation between mean f0 and body size of their subjects, they concluded that "possibly the differences between the two groups relate to crosscultural factors" (p. 119), but that "the reasons for such differences are not clear" (p. 124). Most other studies (e.g., Altenberg and Ferrand, 2006; Hanley *et al*., 1967) have adopted similar measures of mean/median f0 (and on occasion f0 standard deviation) as uncontroversial measures, capable of characterizing f0 range differences across languages.

In light of this background, the aim of this study is to apply the tools developed by Patterson in a comparative study of f0 range in English and German. The choice of these two languages was determined by the fact that there is strong anecdotal evidence that people perceive differences in f0 range between speakers of these languages, with English sounding higher and having more pitch variation than German. British voices (especially female) are often perceived as "over-excited" (Eckert and Laver, 1994) or even "aggressive" (Gibbon, 1998) by German listeners. Conversely, to British listeners, German low-pitched voices may sound "bored" or "unfriendly" (Gibbon, 1998). This belief has even found its way into the German film industry, which—despite a need to match the voices of the dubbing actors to the original ones—goes out of its way to use German dubbing actresses with a lower pitch and/or narrower f0 range than those of original English actresses to avoid this stereotyping (Eckert and Laver, 1994). These differences are reported notwithstanding research which suggests that English and German are intonationally rather similar, albeit with some differences in the phonetic realization of some tonal contrasts (Grabe, 1998).

The principal objective of this study is to apply diverse LTD and linguistic measures in characterizing f0 range for female speakers of German and English in order to identify whether and on what basis or bases the two languages differ. In the process, the validity of the different approaches to measuring f0 range is evaluated, and the findings are discussed in respect of the theoretical questions which they raise and their implications for a number of applied areas where f0 range is an important metric.

II. METHODS

A. Participants

30 female English speakers and 30 female German speakers took part in the study. To avoid effects of gender on f0 range and because of the stronger anecdotal evidence for cross-language f0 differences in females, we limited the study to female speakers only. The English speakers were recruited in Edinburgh, UK, the German speakers in Potsdam, Germany. The target accent for the British group was "Southern Standard British English" (SSBE), for the German group "Northern Standard German" (NSG). The British speakers were selected by advertising for speakers of "Southern English origin," with Southern England defined as the area generally described as showing the BATH-TRAP split (Wells, 1982). While Edinburgh is clearly not in the South of England, it nevertheless has a substantial population of SSBE speakers within the target age range, so recruiting speakers of the target accent in Edinburgh did not pose a problem. The German speakers were selected by an advert posted at the University of Potsdam, asking for speakers of Northern Standard German. Although the literature suggests that for adult speakers of the same sex there is no clear connection between f0 and height or body size more generally (e.g., Kreiman and van Lancker, 2011; Rendall *et al*., 2007), we nevertheless decided to ask subjects to state their body height to control for potential effects of height on specific aspects of f0 range, in the absence of previous work which investigated this aspect of f0. In addition, we assessed the accent of every

speaker by a formal procedure (see below). The English group was between 19 and 38 years of age (mean age 23.2), the German group between 20 and 29 (mean age 22.8).

B. Recording setup

All recordings took place in sound treated rooms. For the Edinburgh recordings a Marantz flash recorder and an AKG condenser microphone was used. The German recordings were performed with a Tascam DAT-recorder and an Audio-Technica condenser microphone. Sampling frequency in both cases was 44.1 KHz.

C. Accent judgement

Alongside the pre-selection of participants for origin, the accent of every participant was assessed. For this purpose, a list of words and short phrases was recorded along with the material for the study proper. For SSBE, this list consisted of an extended version of Wells' (1982) lexical sets; for NSG, this list was composed especially for the study, as a comparable standard set is not available for German. This list focused on vowel quality, vowel length, frication of velar plosives, and final devoicing.

Accent judgements were performed by the second and third author who are native speakers of German and British English, respectively. The exclusion criteria for the English participants were the presence of rhoticity and non-SSBE vowel qualities. For the German participants all pronunciations were compared to the Duden pronunciation standard (Mangold and Grebe, 2005). Exclusion criteria were non-standard vowel qualities, non-standard vowel length and final frication of velar plosives (cf., /tax/ vs /ta:k/ for German <Tag>—"day").

D. Material

Apart from the word list, every speaker also recorded the *Dog and Duck* story (Brown and Docherty, 1995)—which we translated into German for the NSG speakers—and some further material for future studies. The first five sentences of the *Dog and Duck* story were selected for further analysis. The *Dog and Duck* story was chosen as it is a lively text, containing a combination of direct and indirect speech, statements and questions. We expected that this would trigger a range of intonation patterns, and variation in f0 range. Baken and Orlikoff (2000) (p. 172) discuss the relative merits of read versus unscripted speech as a basis for measuring speaking fundamental frequency. They point to small differences in mean f0 across the two speech styles, but note no large differences in "f0 variability." They suggest that since using the same reading passage across participants enables control of a number of possible confounding factors, this is an effective approach to take in cases where reading ability is not at issue.

E. Procedure

1. F0 tracking procedure and artifact correction

F0 tracking (or pitch tracking as it is called in PRAAT) was performed with the PRAAT standard algorithm for f0 tracking, which is based on an autocorrelation method (cf., the PRAAT manual that is integrated in the PRAAT software package). All settings remained at their standard settings, i.e., "pitch floor" was set to 75 Hz, and "pitch ceiling" was set to 600 Hz. For a pitch floor of 75 Hz, PRAAT uses a time step of 10 ms. The authors are aware that PRAAT recommendations for female voices are slightly different, namely 100 Hz for pitch floor and 500 Hz for pitch ceiling. However, as manual correction was part of the process we decided to include a slightly larger frequency range. This potentially meant more laborious manual correction, but allowed us to inspect at least some problematic cases, especially at the lower frequency range, where creak and creaky voice might complicate f0 tracking. We opted for exclusion of creaky voice, as this would be standard procedure for intonation analysis with PRAAT (a pitch floor setting of 100 Hz probably excludes most cases of creaky voice), but the authors are aware that the role of creaky voice in f0 range estimation awaits further empirical research. Based on our method, minima in f0 did not differ greatly across languages (see results section below).

For both the LTD and the linguistic analysis, the f0 contour of every sample was visually inspected in PRAAT (Boersma and Weenink, 2007) and manually corrected for artifacts (e.g., octave errors or spurious f0 values in voiceless parts of the signal).

2. Stylization and labeling

For the linguistic analysis, the sample was labeled for a number of landmarks in the f0 contour (see further below), following Patterson's (2000) approach. Labeling was combined with a stylization procedure, to allow for auditory validation of the intonation labels (see below). We aimed for a simplified representation of the f0 contour that still contained all relevant frequency changes ('t Hart, Collier, and Cohen, 1990).

Labeling and f0 range stylization was performed with "manipulation objects" in PRAAT. These "objects" show the original f0 contour of an utterance, and an additional layer that represents the f0 contour as a succession of editable points, called "pitch points" in PRAAT. The f0 contour is linearly interpolated between the pitch points. Pitch points can be added, deleted, or shifted in time and frequency to modify the contour of an utterance. Original and new contours can be compared directly by auditory and visual analysis.

The manipulation objects were used to derive f0 landmarks via visual, auditory and linguistic inspection. This process was performed in four steps.

(1) All original pitch points of an Intonational Phrase[2] (IP) were deleted, then every IP received an initial and a final pitch point (see Fig. 1).
(2) In the next step, every local f0 maximum and minimum received a pitch point (see Fig. 2). This process was mainly driven by visual inspection, but care was taken to exclude short-term f0 perturbations due to e.g. voiceless plosives.
(3) Additional landmarks were inserted wherever interpolation between the already available landmarks deviated considerably from the original f0 contour (as determined

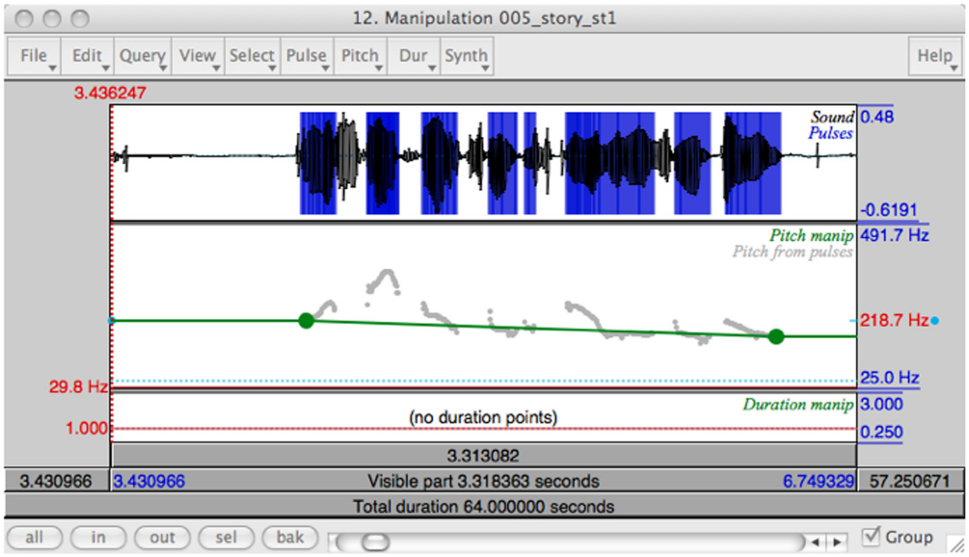


FIG. 1. (Color online) Step 1 of the f0 stylization process which was used, where all original points of an intonational phrase are deleted and replaced by just an initial and final pitch point.

through visual inspection) and auditory comparison between original and new contour did not lead to perceptually equivalent results (see Fig. 3).

(4) In a fourth step, every f0 landmark received a label (see Fig. 4).

Table I describes the labels for all landmarks, and shows stylized representations of the f0 contour environment.

Initial landmarks were only labeled as separate landmarks (transcribed as I) if the initial syllable was unaccented, or—in the case of an initial accent—an initial rise or fall required an additional point to make the stylized contour perceptually identical to the original one. Final landmarks were however always labeled (transcribed as FL or FH).

As a consequence of the stylization process, all phrase-medial landmarks were either local peaks or valleys, or changes in upward or downward slope. Landmarks appearing in stressed syllables (prominence-lending pitch accents) were marked with an asterisk ("starred tone," e.g., H* or L*). Landmarks appearing in unstressed syllables were marked with a single capital letter (e.g., H or L). As such, our labeling system followed principles of the autosegmental-metrical (AM) approach of intonational analysis (as exemplified by Pierrehumbert, 1980). However, our labeling deviates from ToBI style notation (tones and break indices; e.g., Beckman and Ayers Elam, 1997) in a number of ways. Most importantly, our landmarks were not categorized further, so that no assumptions were made about grouping of starred and non-starred single tones into more complex tonal units (e.g., pitch accents with two tones). Moreover, for the purposes of this paper and following Patterson's (2000) approach, we assumed that there is a direct link between local turning points and phonological tones, so that local maxima are assumed to be high tones and local minima are assumed to be low tones. Their status as prominent (H* or L*) or non-prominent (H or L) is solely determined by whether the local peak or valley is realized within a stressed syllable or not.[3]

Tonal accents on downward slopes were transcribed as !H*. For a tonal accent on an upward slope, "$L*" was used. For changes in slope on unaccented syllables, "D" and "U"

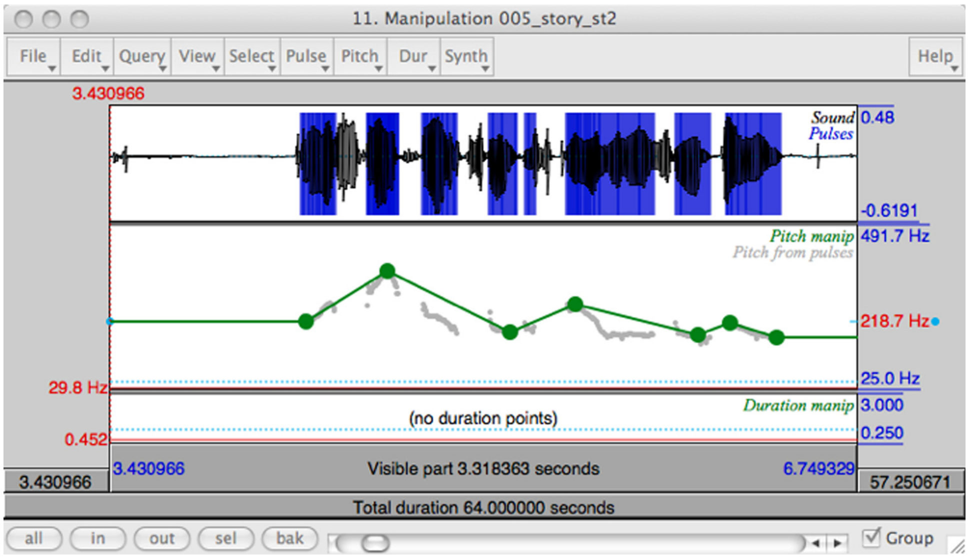


FIG. 2. (Color online) Step 2 of the f0 stylization process which was used, where every local f0 maximum and minimum receives a pitch point.

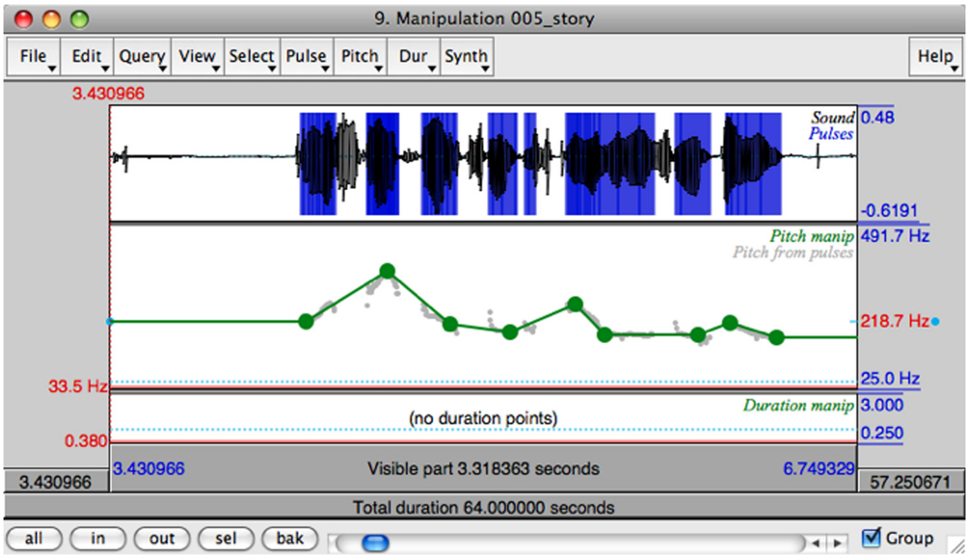


FIG. 3. (Color online) Step 3 of the f0 stylization process, where additional landmarks are inserted wherever interpolation between the already available landmarks deviates considerably from the original f0 contour and auditory comparison between original and new contour does not lead to perceptually equivalent results.

were used. D and U were only marked if there were clear visual and auditory indicators for a change in slope. !H* and $L* were sometimes also marked when there was no change in slope, but auditory criteria suggested the existence of a tonal accent. The landmarks !H*, $L*, D, U, and I were not analyzed further in the present study, as they do not constitute local minima or maxima and were therefore considered of minor importance for f0 range assessments.

Plateaus in the f0 contour were treated as a succession of two landmarks with identical labels, unless the prominence status of the respective syllables was different. In these cases combinations of H* and H or L* and L were used (see Fig. 5).

In order to assess the reliability of our labeling system we calculated inter-rater agreement for a subset of the corpus. A trained labeler independently labeled a random selection of 20% of the English data. We used Cohen's kappa (Cohen, 1960) as the index of inter-rater agreement for the type of landmarks. Average kappa across speakers was 0.67. Agreement at this level is usually deemed as "substantial" (Landis and Koch, 1977), and is at least within the same order of magnitude as inter-rater agreement for various ToBI versions (cf. Breen *et al.*, 2012; Escudero *et al.*, 2012; Yoon *et al.*, 2004). We therefore deemed inter-rater reliability as sufficient for our purposes and proceeded with the labels provided by labeler 1.

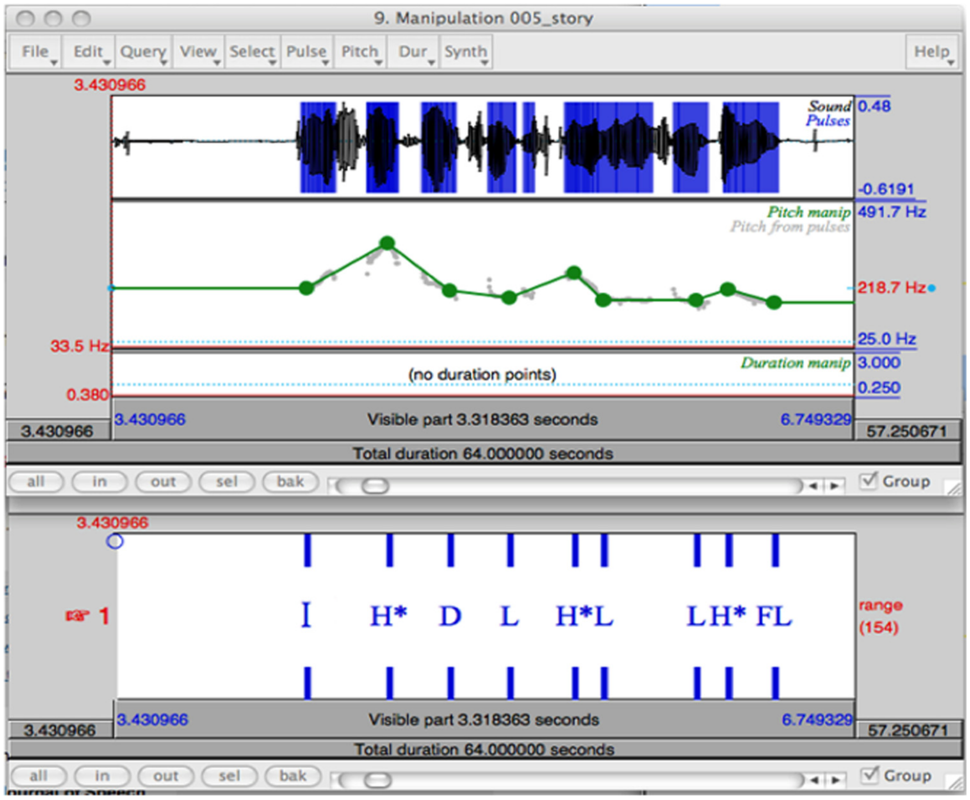
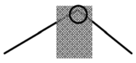
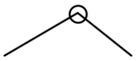
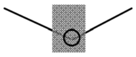
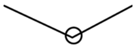
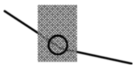
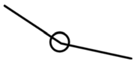
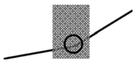
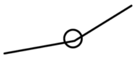


FIG. 4. (Color online) Step 4 of the f0 stylization process which was used, where every f0 landmark receives a label.

TABLE I. Labels used for f0 range analysis. The first column shows the labels used for landmarks in the initial, medial, and final parts of the f0 contour. The second column gives a description of the landmarks. The final column shows stylized environments, where shaded areas mark prominent syllables and circles indicate landmark positions.

Target label	Description	F0 contour
Initial		
I	Phrase-initial value.	
Medial		
H*	Local peak, prominent syllable	
H	Local peak, non-prominent syllable	
H*i/Hi	First H* or H of every IP, mutually exclusive	(same as H*/H)
L*	Local valley, prominent syllable	
L	Local valley, non-prominent syllable	
!H*	Change in downward slope on prominent syllable	
D	Change in downward slope on unaccented syllable	
$L*	Change in upward slope on accented syllable	
U	Change in upward slope on unaccented syllable	
Final		
FH	Final local maximum, higher than the preceding context, or as high as a preceding H or H*	
FL	Final local minimum, lower than the preceding context or as low as a preceding L or L*.	

F0 values for all tonal landmarks were extracted with a PRAAT script. During this process, the initial H* or H of every IP was automatically relabeled as H*i or Hi. In the following, H* and H therefore always denote non-initial peaks, if not indicated otherwise. The LTD measures were extracted over the whole passage for each speaker, using the manually corrected pitch contours. The different linguistic measures were also extracted over the whole passage, after which the speaker-average was taken. Thus, for further processing, every speaker was represented by a single value for each variable.[4] Most of the LTD measures (e.g., speaker mean, median, span, see further below) could be derived by built-in PRAAT functions.[5]

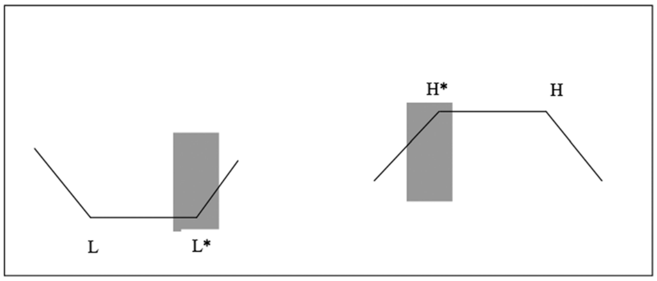


FIG. 5. Examples showing the labeling of low (left) or high (middle) plateaus in the f0 contour.

F. Measures

As mentioned above this study investigated two types of measures across the two languages, in order to assess which of these measures best reflected the f0 level and span differences between the two languages. Table II shows all measures used for language comparison. For LTD level we used mean f0, median, maximum, and minimum f0. For LTD span we used standard deviation (SD), four standard deviations around the mean (SD4), maximum minus minimum f0 (max-min f0), the difference between the 95th and 5th percentile (90% span), the difference between the 90th and 10th percentile (80% span), skew and kurtosis. Our choice of linguistic measures was loosely based on Patterson's (2000) approach (which in turn was based on the AM approach), particularly in that it distinguished between prominent and non-prominent peaks and valleys and separated initial from subsequent peaks, as these were thought to behave differently (at least in English, cf. Patterson, 2000). The linguistic measures we used for level were prominent phrase-initial peaks (H*i), prominent non-initial peaks (H*), initial prominent and non-prominent peaks combined (first peak, i.e., the combined measures of H*i and Hi), non-prominent initial peaks (Hi), non-initial non-prominent peaks (H), prominent valleys (L*), non-prominent valleys (L), and phrase-final lows (FL). Linguistic measures for span comprised the various combinations of H*i-L, H*i-FL, H*-L, H*-FL, first peak–L, and first peak–FL.

All measures were initially made in Hz. To assess the effect of different scales (semitones and ERB), we compared the "best" measures (i.e., those with the largest effect sizes, see below and results section) with a comparable distribution

TABLE II. The LTD and linguistic f0 range measures used for language comparison of level and span. Brackets indicate that for these measures there were too few datapoints for statistical analysis (although descriptive statistics for these measures are reported in the paper).

	LTD	Linguistic
Level	Mean f0	First peak
	Median f0	H*i
	Maximum	H*
	Minimum	(Hi)
		(H)
		(L*)
		L
		FL
Span	SD4	H*i-L
	SD	H*i-FL
	Max-min f0	H*-L
	Max – min ST	H*-FL
	Max-min ERB	H*-FL ST
	90% span	H*-FL ERB
	80% span	First peak–L
	Skew	First peak–L ST
	Kurtosis	First peak–L ERB
		First peak–FL

across the two languages in the LTD and linguistic span group with their transformed counterparts.

ERB and ST transformations were only applied to span measures. ST measures are mainly suitable for frequency differences. If they were to be applied to level measures, an arbitrary reference point would have to be defined. The ERB scale does not require a reference point, but for level measures the transformation is monotone. As we used non-parametric tests for statistical analysis (see below), which are based on ranks, the test outcome would be identical for Hz and ERB.

Finally, in order to establish whether there was a correlation between our f0 range measures and height we calculated a correlation coefficient (Spearman's rho).

G. Statistical analysis

It has been reported that f0 values expressed in Hz are not normally distributed (see, for example, Patterson, 2000). Some researchers have therefore suggested using non-parametric tests for comparison of pitch samples or logarithmic transformations. As we aimed at comparisons of measures on different scales, including Hz and at the same time wanted to apply a common statistical approach for all measures, we used the Shapiro–Wilk test to initially assess whether some of the sample distributions showed significant deviations from normality. As this was indeed the case for a number of variables we decided to apply non-parametric tests in all cases.

Given the large number of dependent variables investigated in this study, a decision had to be made whether to apply a multivariate method or repeated univariate tests (and thus to increase the family-wise error rate). An issue that affects multivariate methods is multicollinearity, i.e., highly correlated dependent variables. For our study we derived a range of measures from the same samples that were conceptually related (e.g., mean, median, quantiles). As some of these were highly correlated we decided against applying a multivariate method. This is also in line with the aim of the study to compare measures of f0 range, instead of building a model of combined effects of different f0 range measures. We therefore decided to apply a series of univariate Mann–Whitney U tests, and adapted for false discovery rate with Benjamini–Hochberg correction (Benjamini and Hochberg, 1995).

As we wanted to compare different measures of f0 range, we analyzed effect size for every variable. Clark-Carter (1997) (p. 455) provides the following formula for effect size for the Mann–Whitney U-test:

$$r = \frac{z}{\sqrt{N}}.$$

This effect size measure was used as a means to quantify which measures best capture the cross-language differences in f0 range of the two languages, interpreting a large ($\geq$0.5) effect size as a very good measure, a medium (0.3–0.49) effect size as a good measure and a small ($<$0.3) effect size as a poor measure (Field, 2005).

III. RESULTS

We refer to Tables III and IV for our results: the means and standard deviations for all measures are reported in Table III; Mann–Whitney U-test results are given in Table IV. A graphical representation of the effect sizes for measures with significant cross-language differences is given in Fig. 6.

Initial inspection of the data revealed that there were many empty cells for some of our measures. For this reason no further Mann–Whitney tests were calculated for the measures affected (i.e., Hi, H, and L*), although we do report the descriptive statistics in Table III.

As expected, the results yielded significant differences in both f0 span and level measures across the groups of German and English females and‘ this was not correlated with their height (with no significant differences found in height between the German and the English group and no significant correlations for height with any of our f0 span or level measures). This cross-language difference is by and large in line with reported stereotypical beliefs of a higher f0 level and wider f0 span for the English females (see Table IV and

TABLE III. F0 range measures used in our study, their N (English/German) and the actual values (with standard deviations in brackets) for English and German female speakers in our study. The values for each measure are given in Hz, unless otherwise specified in the first column.

Measure	N English/German	English	German
LTD-Level			
Mean f0	30/30	216.64 (21.80)	218.04 (14.26)
Median f0	30/30	209.31 (21.77)	213.19 (13.43)
Maximum	30/30	354.79 (60.37)	313.98 (28.69)
Minimum	30/30	154.79 (19.14)	164.26 (11.89)
LTD span			
SD4	30/30	155.48 (43.30)	127.29 (28.43)
SD	30/30	38.87 (10.82)	31.82 (7.11)
Max–min f0	30/30	200.00 (58.19)	149.72 (27.43)
Max–min ST	30/30	14.24 (3.07)	11.20 (1.65)
Max–min ERB	30/30	3.49 (0.82)	2.71 (0.4)
90% span	30/30	122.15 (35.52)	100.58 (20.33)
80% span	30/30	93.60 (27.12)	83.59 (19.15)
Skew	30/30	1.08 (0.41)	0.56 (0.28)
Kurtosis	30/30	4.26 (1.65)	2.70 (0.62)
Linguistic level			
First peak	30/30	278.13 (36.70)	255.98 (22.40)
H*i	30/29	277.97 (37.77)	254.82 (26.64)
H*	29/29	230.47 (23.21)	258.22 (24.56)
(Hi)	12/29	273.02 (42.28)	258.15 (26.24)
(H)	14/30	234.92 (48.48)	249.61 (27.21)
(L*)	17/30	193.51 (32.60)	196.27 (13.64)
L	30/30	193.77 (22.14)	209.47 (17.41)
FL	30/30	174.27 (20.01)	178.14 (11.72)
Linguistic span			
H*i-L	30/29	84.19 (33.50)	45.44 (27.11)
H*i-FL	30/29	103.69 (33.68)	76.66 (25.01)
H*-L	29/29	37.75 (17.24)	50.53 (19.64)
H*-FL	29/29	55.49 (18.88)	79.89 (22.86)
H*-FL ST	29/29	4.78 (1.60)	6.37 (1.59)
H*-FL ERB	29/29	1.09 (0.37)	1.51 (.39)
First peak–L	30/30	84.36 (33.04)	46.51 (20.98)
First peak–L ST	30/30	6.22 (2.26)	3.47 (1.43)
First peak–L ERB	30/30	1.52 (0.56)	0.85 (0.36)
First peak–FL	30/30	103.86 (33.13)	77.84 (19.40)

TABLE IV. Statistics and effect size for Mann–Whitney U-test. An asterisk denotes significance after Benjamini–Hochberg correction (Benjamini and Hochberg, 1995).

Measure	U	Z	*p*-value		Effect size (*r*)
LTD-level					
Mean f0	419.0	−0.458	0.647		0.059
Median f0	381.5	−1.013	0.311		0.131
Maximum	259.0	−2.824	0.005	*	0.365
Minimum	279.5	−2.521	0.012	*	0.325
LTD span					
SD4	277.0	−2.558	0.011	*	0.330
SD	277.0	−2.558	0.011	*	0.330
Max–min f0	203.0	−3.652	0.000	*	0.471
Max–min ST	186.0	−3.905	0.000	*	0.504
Max–min ERB	191.0	−3.829	0.000	*	0.494
90% span	272.0	−2.632	0.008	*	0.340
80% span	350.0	−1.478	0.139		0.191
Skew	136.0	−4.642	0.000	*	0.599
Kurtosis	138.5	−4.611	0.000	*	0.595
Linguistic level					
First peak	285.0	−2.439	0.015	*	0.315
H*i	270.0	−2.502	0.012	*	0.326
H*	176.0	−3.802	0.000	*	0.499
L	263.0	−2.765	0.006	*	0.357
FL	357.0	−1.375	0.169		0.178
Linguistic span					
H*i-L	142.0	−4.442	0.000	*	0.578
H*i-FL	221.0	−3.245	0.001	*	0.422
H*-L	247.0	−2.698	0.007	*	0.354
H*-FL	176.0	−3.802	0.000	*	0.499
H*-FL ST	203.0	−3.382	0.001	*	0.444
H*-FL ERB	189.0	−3.600	0.000	*	0.473
First peak–L	133.0	−4.687	0.000	*	0.605
First peak–L ST	121.0	−4.864	0.000	*	0.628
First peak–L ERB	124.0	−4.820	0.000	*	0.622
First peak–FL	226.0	−3.312	0.001	*	0.428

Fig. 6), but with some notable exceptions as discussed further below.

Of all the span measures that were tested, the ones with the largest effect sizes were the linguistic measure of the average of initial peaks minus the average of non-prominent valleys ("first peak–L," with an effect size of 0.605 when expressed in Hz and 0.628 and 0.622 when expressed on ST and ERB scale, see Table IV), followed by the LTD measures of skew and kurtosis (with effect sizes of 0.599 and 0.595, respectively). We found a positively skewed f0 distribution for both languages (having an asymmetric distribution with a longer right tail, but with higher skew values for English than for German speakers). Similarly, English speakers also had a higher kurtosis than our German speakers. Most of the German kurtosis values are below 0, indicating a distribution that has a flatter top and thinner tails than a normal distribution. Most English values are above 0, indicating pointier tops and fatter tails. For our data, skewness and kurtosis are highly correlated (Spearman's $\rho = 0.914$, $p < 0.001$), suggesting that the values are influenced by a common factor. All other span measures that were calculated yielded medium to large effect sizes except 80% span, which was the only LTD span measure that did not show significant cross-language differences.

The largest effect size for the level measures was found for the prominent non-initial peaks (H*, with a large effect size of 0.499, albeit in a different direction than the one that prevailed overall, as discussed below), followed by maximum f0 (with a medium effect size of 0.365). Note that mean f0 and median f0 (measures which are most commonly employed to investigate cross-language differences) turned out not to be significantly different across the two languages, nor were the final low measures (although a trend was observed for final low measures to be lower in English speakers, see Table III).

In general, we found that average effect sizes were larger for span (0.469) than for level measures (0.284), and for linguistic (0.449) than for LTD measures (0.364). Only minor differences were found in the effect sizes of the different scales used for the span measures, with marginally larger effect sizes for the span measures that were expressed on a ST or ERB scale compared to the corresponding Hz measures.

Importantly, some of the linguistic landmarks which were measured showed a difference between the two languages opposite to that which prevailed overall. Figure 7 shows the average values in Hz for each landmark by German and English speakers. For example, while initial prominent peaks (H*i) in English are significantly higher than in German, non-initial peaks (H*, H) showed the reverse effect, as did non-prominent valleys (L). This influenced the differences in span: at the beginning of intonational phrases, f0 span for English females was wider than that for German females, but the reverse was true for the later parts of intonational phrases (note that the landmarks in Fig. 7 are shown in an order that approximates their order in an IP). Visual inspection of Fig. 7 suggests that the combination of higher initial prominent peaks and lower non-initial prominent peaks in English females may lead to a more declining (or downstepping) intonational phrase as compared to German females where prominent peaks appear relatively similar across initial and non-initial positions in the phrase. We decided to test whether English and German females realize the differences between initial and non-initial prominent peaks differently, by using Wilcoxon signed-ranks tests. These confirmed that there is indeed a highly significant difference between initial and non-initial prominent peaks (H*i versus H*) in English ($Z = -4.681$, $p < 0.001$) but not in German ($Z = -0.512$, $p = 0.608$) female speakers. In other words, the realization of prominent peaks varies depending on their position (initial or non-initial) in the intonational phrase in English but not in German females. This, combined with the fact that final landmarks (FL) are not significantly different (compare Tables III and IV) and values for the prominent valleys (L*) appear similar between the groups[6] seems to point towards a flatter intonation contour in German than in English female speakers.

A further outcome of the analysis was the finding that the comparative frequency of the landmarks found in the f0 contours varies across the two languages. Figure 8 shows the distribution of landmarks on prominent syllables used by the German and English females in our study. As can be seen,

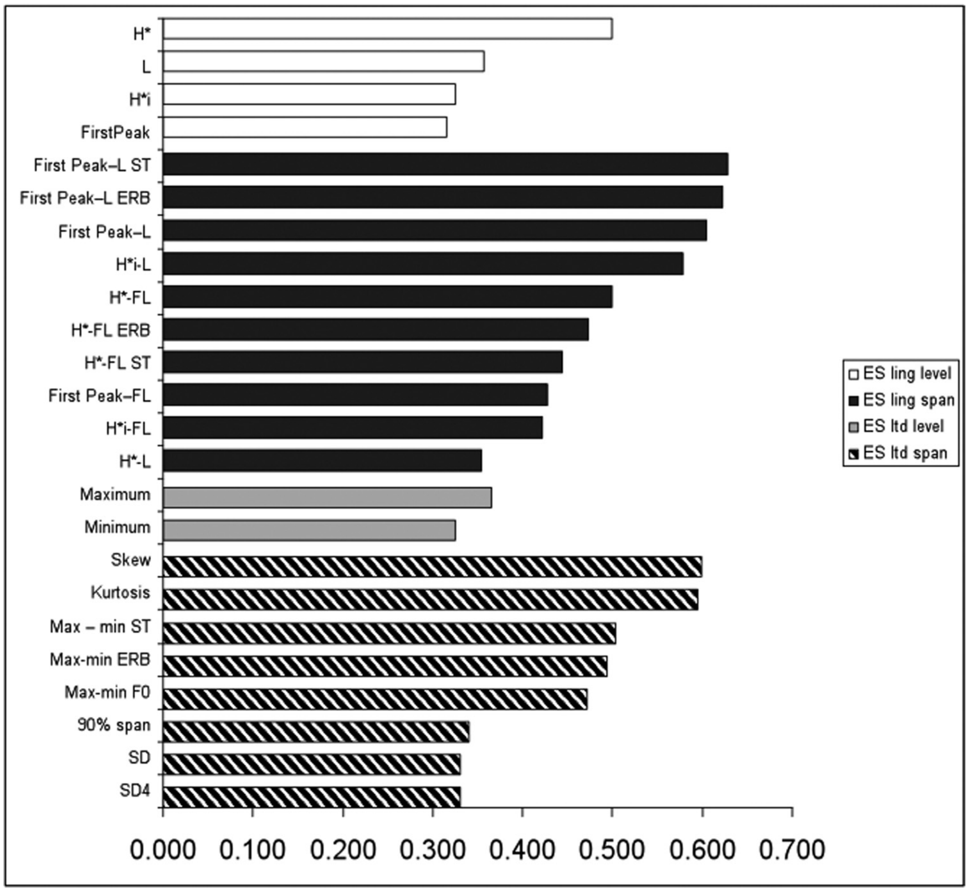


FIG. 6. Graphical representation of the effect sizes (ES) for all linguistic level measures (white bars), linguistic span measures (black bars), LTD level measures (gray bars), and LTD span measures (striped bars). Effect sizes are largest for span measures, and larger for linguistic than LTD measures.

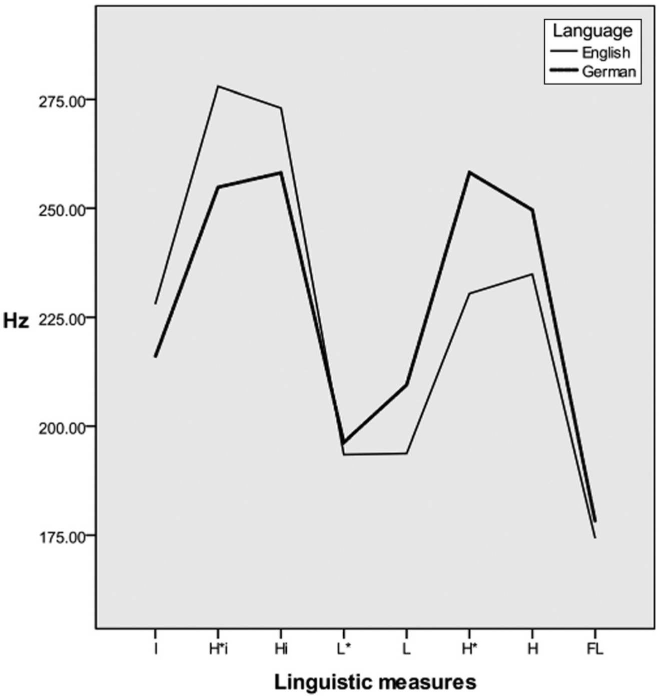


FIG. 7. Realization in Hz of the linguistic measures by German and English females in our study. The thin line represents English speakers, the thick line German speakers. The linguistic measures presented here are phrase-initial f0 (I), initial peaks on prominent syllable (H*i), non-prominent initial peaks (Hi), valleys on prominent syllable (L*), non-prominent valleys (L), prominent non-initial peaks (H*), non-initial non-prominent peaks (H), and phrase-final lows (FL). Differences in the direction between the two languages can be observed for some linguistic measures.

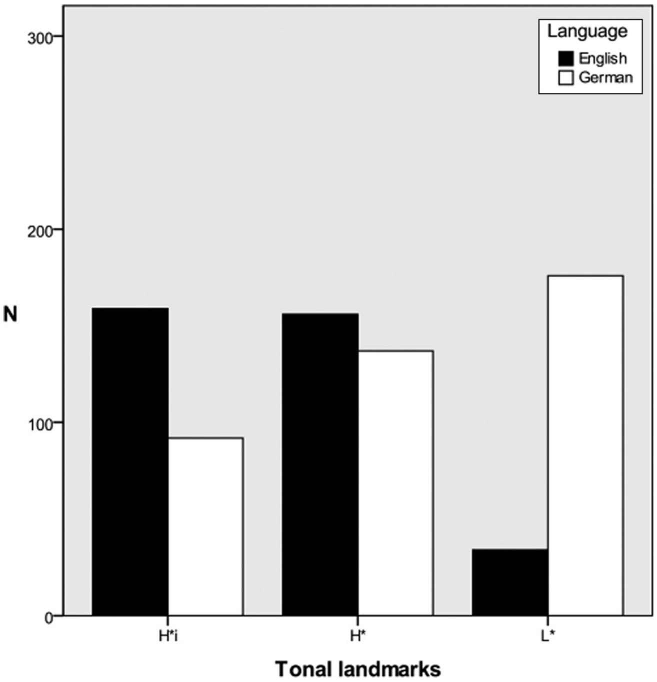


FIG. 8. Distribution of landmarks on prominent syllables produced by the German (dark bars) and English speakers (light bars) in our study. English females produce more instances of prominent phrase-initial peaks (H*i) and prominent non-initial accent peaks (H*) than German speakers, whereas the opposite holds true for the prominent valleys (L*).

English speakers used considerably more initial (H*i) and non-initial prominent peaks (H*) than German speakers; German speakers used more prominent valleys (L*) compared to English speakers. This indicates that the intonational structure of English may lead speakers to use the upper end of their range more than the lower end, particularly at the start of a phrase, whereas the reverse pattern may be more common for the German speakers.

Figure 9 gives a visual representation of the span and level for all English and German speakers of our study, by plotting two representative measures illustrating the differences in span and level across the two groups of speakers in line with reported beliefs. This figure illustrates that the majority of German speakers cluster at the lower end of the x-axis (representing span), whereas English speakers cluster mostly at the higher end of the x-axis. Similarly, more clustering can be observed at the lower end of the y-axis (representing level) for the German speakers, whereas higher values can be observed for the English speakers. This figure also shows that the cross-language differences in f0 are not necessarily present in all speakers (see, for example, speaker 15), but are indeed present collectively (see discussion for further comment on this).

IV. DISCUSSION

The results of our study add to the growing number of studies reporting that speakers of different languages or dialects may use characteristically different f0 ranges and provide validation of the anecdotal belief that female speakers of English and German differ in f0 range. Our findings show that these differences occur along two dimensions, span and level, and are captured by measures that make reference to linguistically relevant landmarks as well as by more global non-linguistic measures. Yet, although both types of measures identify cross-language differences in f0 range, they may not be equally effective or informative. While LTD measures have the advantage of relative ease of computation, the linguistic measures were found to be better predictors of language membership with larger effect sizes than those obtained for the majority of LTD measures of which only skew and kurtosis reach comparable effect sizes. More importantly, LTD measures fail to capture important information as to the underlying source of cross-language differences. Our results show that f0 range is influenced by differences between the two languages that are linguistic in nature. In particular, the linguistic measures used in this study highlighted differences in the realization of tones at different points in the intonation contour, alongside some differences in the typical frequency of distribution of tones. That is, our results show that f0 range is influenced by the phonological and/or phonetic conventions of the language being spoken and is not solely an artifact of physiological factors or cultural differences, as often assumed (e.g., Altenberg and Ferrand, 2006; Dolson, 1994; Hanley, Snidecor, and Ringel, 1967; Keating and Kuo, 2010).

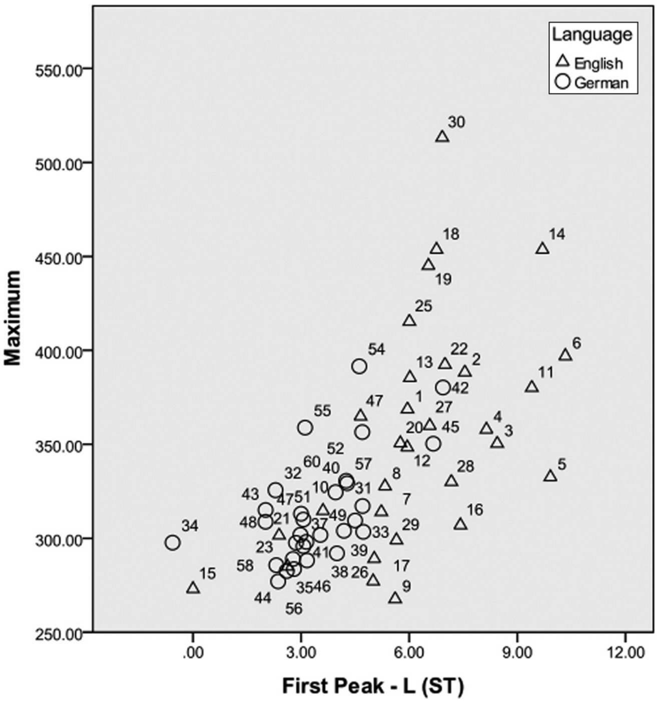


FIG. 9. Scattergraph illustrating representative measures for span and level showing differences in English and German in the female speakers of the study. Triangles represent German speakers, circles represent English speakers. The numbers represent the individual speakers.

This raises a number of issues. First, the finding that some landmarks show a difference between the two languages opposite to stereotypical beliefs and prevailing findings—particularly towards the later parts of intonation phrases—suggests that perceptual evaluation by listeners is not evenly influenced by various parts of the intonation phrase. It may, for instance, be the case that listeners base their judgments on the very beginning of the phrase (which is higher in English than in German speakers), on the overall contour shape of the speakers (which is flatter for German speakers), or on particular linguistic landmarks (e.g., H*i). However, it remains to be determined in a perceptual investigation which of the f0 range parameters and which point in the intonation phrase underpin the cross-language differences that people perceive, and what the importance of global versus local f0 range influences are on such perception (see, e.g., Gussenhoven, Repp, Rietveld, Rump, and Terken, 1997; Pierrehumbert, 1979; Terken, 1994).

A second issue that remains to be explored is that of individual differences. Although our results clearly show that the two groups of speakers were statistically differentiated, there was nevertheless a degree of overlap in the range deployed by many speakers of both languages. That is, the f0 range difference is a characteristic of the collective, but not necessarily of individual speakers. Further research is needed to establish what it means for some speakers' f0 range not to fall within the characteristic range of their linguistic community.

Our findings also help clarify an important non-lexical aspect of communication which might be beneficial for clinical and second language (L2) acquisition research and practice. Most investigations of f0 range in clinical populations are based on measures of f0 SD and mean or median f0. While such measures may be capable of identifying f0 range differences between clinical and healthy populations, they do not give a sense of the possible underlying cause of the problem. Such global LTD measures are not capable of establishing whether observed differences are due to the f0 range realization (perhaps caused by a lack of control at the motor execution level) or intonational structure (possibly at the higher processing level) being affected. Linguistic measures of f0 range have the potential to differentiate between these different underlying causes. This suggests that linguistic measures potentially have great value for the investigation and management of disordered speech, although the challenge is to translate these measures into tools that are practical and less time-consuming for clinical use. Similarly, linguistic measures may also give more insight into the reason for the often reported difficulty of L2 learners in adopting f0 ranges that are appropriate for the target language. Our results show that the cross-language differences in f0 range arise from differences in intonational structure (along with the typical frequency of distribution and realization of tones). It is therefore likely that L2 learners are not simply transferring the phonetic routines of their native language but that their reliance on L1 intonation patterns may lie at the heart of the problem. Further research is needed to establish whether this is indeed the case.

Finally, it is clear from the results of our study that there are certain gains to be made by the use of the linguistic measures in cross-language comparisons of f0 range. Further research needs to establish the full extent of these gains. In particular, so far we have only applied this method to an investigation of f0 range in female speakers in a read speech sample. It will take further experimentation to establish whether the findings extend to male speakers, other languages (e.g., it remains an empirical question whether and how this method can be used for comparisons of more dissimilar intonation languages or even tone languages), and other factors that might influence f0 range. For example, it might prove useful to investigate the effect of different speech styles on f0 range, such as that found by Keating and Kuo (2010). Most importantly, the present study used a method of comparing effect sizes to establish which measure best characterizes cross-language differences in f0 range. It is crucial to follow this up with perceptual experimentation in order to determine whether the measures with large effect sizes correspond to the cross-language differences which listeners perceive. Further perception experiments can determine which measures of f0 span and f0 level and which measurement scales correlate with listeners' perceptual sensitivity to the observed cross-language differences.

ACKNOWLEDGMENTS

This research was supported by the Economic and Social Research Council (ESRC) through Grant No. RES-000-22-1858. Additional funding was provided by the Arts and Humanities Research Council (AHRC) through Fellowship AH/J000302/1 to the first author. This funding is gratefully acknowledged. We thank Frank Kuegler and his team for his valuable support with the data collection. The authors also wish to thank the German and English speakers who participated in this study, and Dr. Catherine Dickie for assistance with the data. Finally, we thank Professor D. R. Ladd and an anonymous reviewer for their valuable comments on earlier versions of the paper.

[1] This thesis is available for download via the British Library's Electronic Theses Online Service (ETHOS), http://ethos.bl.uk/OrderDetails.do?did=1anduin=uk.bl.ethos.492996 (Last viewed December 28, 2011). After registration with this service, individuals (whatever their location) can download a free electronic version of the thesis.

[2] Intonational phrases were identified auditorily as a perceived disjuncture in f0 contours. The disjuncture consisted in most cases of silence (i.e., a pause), lengthening of the last syllable before the end of a phrase, a melodic feature (i.e., a high or low boundary tone), or a combination of these (e.g., Beckman and Ayers Elam, 1997).

[3] Our decision to assume a direct relationship between turning points and phonological tones was driven by practical reasons so as to ensure consistency in our labeling. However, tones and turning points may not necessarily map in a one-to-one fashion, so that some tones may not be realized as turning points and some turning points may not constitute an underlying phonological tone (see, e.g., Ladd, 2008, for a discussion).

[4] For some linguistic measures there were missing values, as some speakers never produced certain landmarks.

[5] Skew and kurtosis measures had to be implemented separately. We used the functions given in the NIST/SEMATECH e-Handbook of Statistical Methods (2010) (Last viewed October 26, 2010). For kurtosis we chose the formula that calculates a kurtosis of 0 for the normal distribution.

[6] Note that we were unable to statistically test cross-language differences in L* as there were too many missing values. However, given that the means across the groups are less than 4 Hz apart, it seems reasonable to say that they appear relatively similar.

Altenberg, E. P., and Ferrand, C. T. (**2006**). "Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English-Cantonese young adult women," J. Voice **20**(1), 89–96.

Awan, S. N., and Mueller, P. B. (**1996**). "Speaking fundamental frequency characteristics of white, African American, and Hispanic kindergartners," J. Speech. Hear. Res. **39**(3), 573–577.

Baken, R. J., and Orlikoff, R. F. (**2000**). *Clinical Measurement of Speech and Voice*, 2nd ed. (Singular Publishing Group, San Diego, CA).

Banse, R., and Scherer, K. R. (**1996**). "Acoustic profiles in vocal emotion expression," J. Pers. Soc. Psychol. **70**(3), 614–636.

Beckman, M., and Ayers Elam, G. (**1997**). *Guidelines for ToBI Labeling*, version 3 (Ohio State University, Ohio).

Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the false discovery rate—a practical and powerful approach to multiple testing," J. R. Statist. Soc. B **57**(1), 289–300.

Boersma, P., and Weenink, D. (**2007**). "Praat: Doing phonetics by computer (version 4.6) [computer program]," http://www.praat.org/ (Last viewed May 14, 2007).

Breen, M., Dilley, L. C., Kraemer, J., and Gibson, E. (**2012**). "Inter-transcriber agreement for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)," Corpus Linguist. Linguist. Theory (in press).

Brown, A., and Docherty, G. J. (**1995**). "Phonetic variation in dysarthric speech as a function of sampling task," Eur. J. Disord. Commun. **30**(1), 17–35.

Chen, S. H. (**2005**). "The effects of tones on speaking frequency and intensity ranges in Mandarin and Min dialects," J. Acoust. Soc. Am. **117**(5), 3225–3230.

Clark-Carter, D. (**1997**). *Doing Quantitative Psychological Research: From Design to Report* (Psychology Press, Hove, East Sussex).

Cohen, J. (**1960**). "A coefficient for agreement for nominal scales," Educ. Psychol. Meas. **20**, 37–46.

Deutsch, D., Le, J., Shen, J., and Henthorn, T. (**2009**). "The pitch levels of female speech in two Chinese villages," J. Acoust. Soc. Am. **125**(5), EL208–EL213.

Diehl, J. J., Watson, D., Bennetto, L., Mcdonough, J., and Gunlogson, C. (**2009**). "An acoustic analysis of prosody in high-functioning autism," Appl. Psycholinguist. **30**(3), 385–404.

Dilley, L. C., and Brown, M. (**2007**). "Effects of pitch range variation on f0 extrema in an imitation task," J. Phonetics **35**(4), 523–551.

Dolson, M. (**1994**). "The pitch of speech as a function of linguistic community," Music. Percept. **11**(3), 321–331.

Eady, S. J. (**1982**). "Differences in the F0 patterns of speech: Tone language versus stress language," Lang. Speech **25**, 29-42.

Eckert, H., and Laver, J. (**1994**). *Menschen und ihre Stimmen: Aspekte der vokalen Kommunikation (Humans and their Voices: Aspects of Vocal Communication)* (Psychologie Verlags Union, Weinheim).

Escudero, D., Aguilar, L., Vanrell, M. M., and Prieto, P. (**2012**). "Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labelling system," Speech Communications, retrieved from http://prosodia.upf.edu/home/arxiu/publicacions/escudero-et-al_analysis-intertranscriber-consistency-cattobi.pdf (Last viewed December 21, 2011).

Field, A. (**2005**). *Discovering Statistics using SPSS*, 2nd ed. (SAGE Publications, London).

Gibbon, D. (**1998**). "German Intonation," in *Intonation Systems: A Survey of Twenty Languages*, edited by D. J. Hirst and A. Di Christo (Cambridge University Press, Cambridge, MA), pp. 78–95.

Grabe, E. (**1998**). "Comparative intonational phonology: English and German," Ph.D. thesis, Max Planck Institute for Psycholinguistics Nijmegen, Max Planck Institute Series in Psycholinguistics No. 7, Wageningen, Ponsen en Looien.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H., and Terken, J. (**1997**). "The perceptual prominence of fundamental frequency peaks," J. Acoust. Soc. Am. **102**(5), 3009–3022.

Hanley, T. D., Snidecor, J. C., and Ringel, R. L. (**1967**). "Some acoustic differences among languages," Phonetica **14**, 97–107.

Hirschberg, J., and Ward, G. (**1992**). "The influence of pitch range, duration, amplitude, and spectral features on the interpretation of the rise fall rise intonation contour in English," J. Phonetics **20**(2), 241–251.

Hollien, H., Hollien, P. A., and de Jong, G. (**1997**). "Effects of three parameters on speaking fundamental frequency," J. Acoust. Soc. Am. **102**(5), 2984–2992.

Hubbard, K., and Trauner, D. A. (**2007**). "Intonation and emotion in autistic spectrum disorders," J. Psycholinguist. Res. **36**(2), 159–173.

Keating, P., and Kuo, G. (**2010**). "Comparison of speaking fundamental frequency in English and Mandarin," UCLA Work. Papers Phonetics **108**, 164–187.

Kreiman, J., and Van Lancker Sidtis, D. (**2011**). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (John Wiley and Sons, Chichester).

Ladd, D. R. (**2008**). *Intonational Phonology*, 2nd ed. (Cambridge University Press, Cambridge).

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (**1985**). "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," J. Acoust. Soc. Am. **78**(2), 435–444.

Landis, J., and Koch, G. (**1977**). "The measurement of observer agreement for categorical data," Biometrics **33**(1), 159–174.

Liberman, M., and Pierrehumbert, J. (**1984**). "Intonational invariance under changes in pitch range and length," in *Language Sound Structure*, edited by M. Aronoff, R. Oehrle, F. Kelley, and B. W. Stephens (MIT Press, Cambridge, MA), pp. 157–233.

Majewski, W., Hollien, H., and Zalewski, J. (**1972**). "Speaking fundamental frequency of Polish adult males," Phonetica **25**(2), 119–125.

Mangold, M., and Grebe, P. (**2005**). *Duden Aussprachewörterbuch (Duden Pronunciation Dictionary)*, 6th ed. (Dudenverlag, Mannheim).

Nishio, M., and Niimi, S. (**2008**). "Changes in speaking fundamental frequency characteristics with aging," Folia Phoniatr. Logo. **60**(3), 120–127.

NIST/SEMATECH e-Handbook of Statistical Methods, (**2010**). http://www.itl.nist.gov/div898/handbook/ (Last viewed October 26, 2010).

Patterson, D. (**2000**). "A linguistic approach to pitch range modelling," Ph.D. thesis, University of Edinburgh, Edinburgh.

Pierrehumbert, J. (**1979**). "Perception of fundamental-frequency declination," J. Acoust. Soc. Am. **66**(2), 363–369.

Pierrehumbert, J. (**1980**). "The phonology and phonetics of English intonation," Ph.D. thesis, MIT, Cambridge, MA.

Rendall, D., Vokey, J. R., and Nemeth, C. (**2007**). " Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size," J. Exp. Psychol. Hum. Percept. Perform. **33**(5), 1208–1219.

Sobin, C., and Alpert, M. (**1999**). "Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy," J. Psycholinguist. Res. **28**(4), 347–365.

Terken, J. (**1994**). "Fundamental-frequency and perceived prominence of accented syllables II: Nonfinal accents," J. Acoust. Soc. Am. **95**(6), 3662–3665.

't Hart, J., Collier, R., and Cohen, A. (**1990**). *A Perceptual Study of Intonation* (Cambridge University Press, Cambridge).

Van Bezooijen, R. (**1995**). "Sociocultural aspects of pitch differences between Japanese and Dutch women," Lang. Speech **38**, 253–265.

Van Dommelen, W. A., and Moxness, B. H. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," Lang. Speech **38**, 267–287.

Wells, J. C. (**1982**). *Accents of English* (Cambridge University Press, Cambridge), Vols. 1-3.

Yoon, T., Chavarria, S., Cole, J., and Hasegawa, M. (**2004**). "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," Proc. Interspeech **2004**, 2729–2732.