DOI: 10.17469/O2108AISV000009

CAROLINA LINS MACHADO

A cross-linguistic study of between-speaker variability in intensity dynamics in L1 and L2 spontaneous speech

Dynamic aspects of the amplitude envelope appear to reflect speaker-specific information. Intensity dynamics characterized as the temporal displacement of acoustic energy associated to articulatory mouth opening (positive) and closing (negative) gestures was able to explain between-speaker variability in read productions of native speakers of Zürich German. This study examines positive and negative intensity dynamics in spontaneous speech produced by Dutch speakers using their native language and English. Acoustic analysis of informal monologues was performed to examine between-speaker variability. Negative dynamics explained a larger quantity of inter-speaker variability, strengthening the idea of a lesser prosodic control over the mouth closing movement. Furthermore, there was a significant effect of language on intensity dynamics. These findings suggest that speaker-specific information may still be embedded in these time-bound measures despite the language in use.

Keywords: inter-speaker, cross-linguistic, variability, intensity, dynamics.

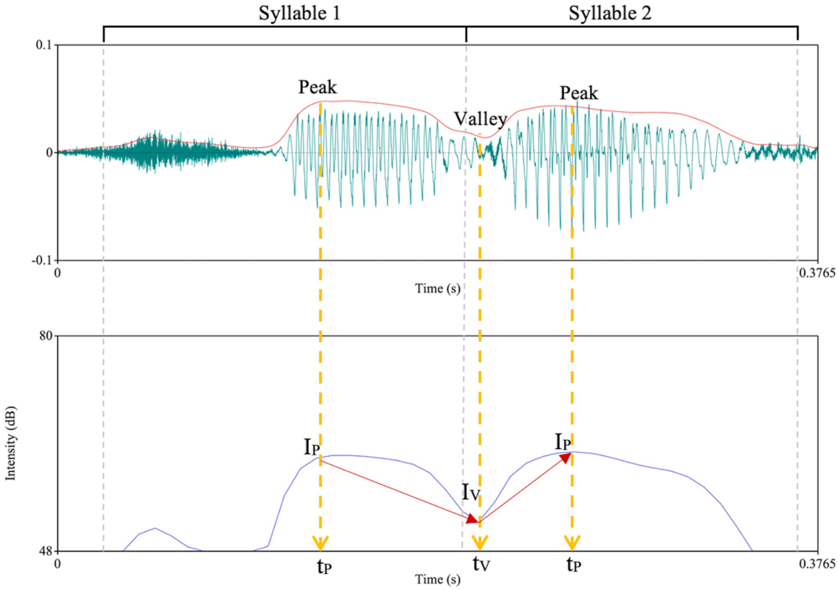
1. *Introduction*

Speech is a dynamic process involving the articulators in our vocal tract that give rise to phonetic sounds. These sounds carry meanings that are specific to each language, making possible the communication of people sharing the same linguistic background. Other than meaning, speech also contains information related to the speaker (Coulthard, Johnson & Wright, 2016: 136). This extra-linguistic information, considered to be a by-product of speaker-specific biomechanical characteristics (Perrier, Winkler, 2015), is extremely valuable in the field of forensic phonetics, where, among other tasks, speaker comparison is often employed.

Speaker comparisons involve comparing speech samples of an unknown speaker to samples of a known speaker to determine whether the unknown samples may belong to the known speaker or to a different speaker. This task can be described as an auditory-acoustic analysis where an experienced forensic phonetician performs an aural-perceptual investigation and examines the acoustic features of the speech signal (Rose, 2002; Coulthard et al., 2016). Among these acoustic features, measures of fundamental frequency (Rose, 2002; Gold, French, 2011) and vowel formants (Goldstein, 1976; McDougall, 2007; He, Zhang & Dellwo, 2019) are some of the most studied parameters, which have been extensively employed in speaker comparisons. However, there are still understudied acoustic features containing significant speaker-specific information, one of such being intensity dynamics.

Intensity by itself is not considered a useful discriminative feature (other than in the context of formants) because of how easily it can be distorted (Hollien, 1990: 198). However, considering the temporal organization of intensity provides a different approach to the use of this acoustic feature. Intensity dynamics is an aspect of speech rhythm, which has been considered a useful parameter for speaker discrimination (Gold, French, 2011: 302). Therefore, investigating inter-speaker variation in intensity dynamics may further contribute to understanding this rhythmic aspect of speech and provide insight into whether it could be useful in forensic applications.

Figure 1 - *The upper plot contains an oscillogram of the Dutch word "student" in teal and its amplitude envelope (superimposed in red). The lower plot illustrates the intensity curve, its peak (I_P) and valley (I_V) values, and time points associated with them (t_P and t_V). These plots are based on the description of positive and negative dynamics by He, Dellwo (2017)*



Intensity dynamics can be understood as the rate of energy increase and decrease in the acoustic signal and is calculated by analyzing the amplitude envelope of the acoustic signal (He, Dellwo, 2017). This concept is demonstrated in Figure 1. In the disyllabic Dutch word "student" there are two syllabic peaks, i.e. places with large amounts of energy in a syllable, and one place with relative low amounts of energy between them, i.e. a valley. I_P are peak points where intensity reaches its maximum value, and I_V is the point where intensity is at its minimal relative to these peaks. Positive dynamics is the rate of increase in intensity from a valley (I_V) to its right adjacent peak (I_P). Negative dynamics is rate of decrease in intensity between a peak (I_P) and the next valley point in time (I_V). In the figure above, negative dynamics is demonstrated in the intensity curve (lower plot) by the red secant line $\overrightarrow{I_P I_V}$ and positive dynamics by the secant line $\overrightarrow{I_V I_P}$.

Changes in intensity are mainly a result of subglottal pressure variation, although both glottal and supra-glottal regions also affect intensity (Fry, 1979; Stevens, 2000). Regarding the supra-glottal region, it has been proposed and empirically demonstrated that the size of mouth opening is one of the factors determining the overall intensity of the speech stream (Summerfield, 1992; Chandrasekaran, Friston, Trubanova, Stillittano, Caplier & Ghazanfar, 2009; Titze, Palaparthi, 2018). Chandrasekaran et al. (2009) found evidence for the relationship between intensity and the articulatory movements responsible for mouth opening and closing gestures. This study established that the amplitude envelope is closely related to the time course of the opening and closing mouth gestures in both read and spontaneous speech samples in two languages, English and French. Furthermore, the authors observed a significant amount of intra- and inter-speaker variability in the temporal patterns of both mouth gestures and the amplitude envelope (Chandrasekaran et al., 2009: 5). Their results were later supported by He, Dellwo (2017), who suggested that inter-speaker variability in the temporal organization of intensity contours of Zürich German speakers may reveal the influence of speaker-specific neurophysiological characteristic over mouth opening and closing movements.

Speaker-specific effects on dynamic acoustic features reflect behavioral variation (Kitamura, Akagi, 2007), i.e. the idiosyncratic way a person operates their articulators to produce speech. Speech articulation is so particular to a speaker that even twins, who have the same anatomical structures allowing them to produce the same canonical phonetic segments, show variation in their production of acoustic dynamic features (Zuo, Mok, 2015). Interestingly, He, Dellwo (2017) observed that in read speech productions negative intensity dynamics showed more between-speaker variability than positive dynamics. They interpreted the lesser variability in positive dynamics as a result of greater prosodic control over the mouth opening gesture during speech production, suggesting that this gesture may exhibit less speaker-specific information due to its function. The authors argued that positive dynamics ask for a more controlled mouth opening to reach the presumed articulatory state of a phonetic segment (phonetic target). Mouth closing gestures on the other hand, were believed to be realized under less prosodic control. That is, once the phonetic target has been reached, speakers may reduce control over this articulatory gesture, which in turn can result in movements exhibiting speakers' behavioral and biological characteristics.

The same result was demonstrated in an earlier study, where De Nil, Abbs (1991) found a wide variety of mouth closing sequencies involving the lower lip, upper lip and jaw in the production of the same utterance by different speakers. Besides the different closing sequencies, they also noticed that some patterns were used more frequently by some speakers than by others. The fact that each of these articulators present a particular morphological structure (Perrier, Winkler, 2015) and are employed differently between-speakers helps understanding the variation found in He, Dellwo's (2017) results. Overall, their study offered a significant contribution to forensic phonetics, providing, as previously mentioned, an additional facet of speech rhythm. However, although a seemingly promising parameter, inter-speaker differences in intensity dynamics should also be studied under other conditions for a better understanding of this feature. Therefore, generalizations and replications in other languages and speech styles are necessary.

Speech production of native and non-native languages show similarities and differences. The underlying mechanisms of speech production in first (L1) and second (L2) languages share similarities related mainly to the mechanical apparatus used during speech production, which is theoretically the same for every healthy speaker (Hixon, Weismer & Hoit, 2020; Marchal, 2009). Differences in L1 and L2 productions are found in other complex mechanical and cognitive actions taking place before and during speech articulation, which are believed to be influenced by language-specific characteristics (Flege, 1995; Best, Tyler, 2006; Escudero, 2009). Although speech is perceived as a highly automatic undertaking in the L1, this is far from true, since, before words come out of our mouths, an utterance needs to be planned and structured (Levelt, 1989). The same is true for the L2, with added constraints related to the speaker's knowledge of this language, and effects stemming, for instance, from the L1 phonology (Kormos, 2006).

Similar to differences owed to language specific constraints are differences across speaking styles. It has been proposed that spontaneous speech contains exclusive phonetic patterns setting it apart from other styles, such as read speech (Simpson, 2013). These patterns may be a result of the communicative situation a speaker is in, involving different factors that influence speech articulation and the resulting acoustic signal (Simpson, 2013: 163). For example, while reading a passage, speakers tend to focus more on the vocalization of the utterance, since there is no need to formulate the message being delivered because it is already given in the text. Contrariwise, when people are talking, information is in the foreground and attentional resources are being divided between the formulation of the content and the act of vocalization.

Together, differences between languages and speaking styles along with the singular way speakers use their anatomically distinctive speech apparatus have a significant effect on the acoustic features of speech. Therefore, this study seeks to fill the gap in our current knowledge by investigating whether intensity dynamics also vary between speakers in spontaneous productions in their L1 and in an L2, and whether there is an influence of language over this feature. More specifically, this study seeks to answer the following questions:

- *RQ1*: Do measures of intensity dynamics vary between native Dutch speakers?
- *RQ2*: Is this between-speaker variability also present when these speakers use English, a second language they are proficient in?
- *RQ3*: Does language influence intensity dynamics?

Regarding *RQ1* and *RQ2* I hypothesize that between-speaker variability in intensity dynamics will be evident despite the language spoken by the speaker, since this variability is a result of speaker-specific characteristics (He, Dellwo, 2017), meaning that productions of native Dutch speakers will vary between speakers in the L1 and in the L2 English. Regarding RQ3, I hypothesize that there will be an effect of language on intensity dynamics, analogous to previous studies in speech dynamics (Schwartz, Kaźmierski, 2019).

2. *Methodology*

2.1 Corpus

The D-LUCEA corpus was used in this study (see Orr, Quené, 2017 for further details), from which 51 female native Dutch speakers (age range of approx. 17–26 years with no reported speech and hearing disorders) were selected based on the quality of their recording. Participants were recruited at University College Utrecht (UCU) and they reported their proficiency in English by providing the results of a formal proficiency exam, which is an entry requirement for this language at UCU with the minimum level of proficiency similar to B1 according to the Common European Framework of Reference for Languages. Additionally, information regarding their language background was collected via a questionnaire where speakers had to report their age of acquisition and degree of exposure to the language (Orr, Quené, Beek, Diefenbach, Leeuwen & Huijbregts, 2011). Each participant was simultaneously recorded via eight microphones in a quiet furnished office with at least one facilitator seated at the opposite side from the speaker (Orr, Quené, 2017). For this study the selected recordings were the ones captured by the microphone closest to the speaker (Sennheiser Headset HSP 2ew; 44.1 kHz; 16 bit), since this microphone had little variation in the distance between the microphone and the speaker's mouth.

From a total of six performed speaking tasks, two two-minute-long prepared informal monologues on a free topic in English (L2) and in Dutch (L1) were selected for this study. Most speakers repeated the same monologue in both languages; however, some of them simply continued the monologue started in one language. These monologues were manually annotated by two annotators and checked by a third annotator at four levels: Language spoken, speech type, speech and silence intervals, and an orthographic transcription of the utterances. Additionally, two more levels were annotated by the author; namely, stretches of fluent uninterrupted speech, which were manually selected to ensure precision, followed by an automatic segmentation of these stretches into smaller chunks. The nature of these chunks is described in the following section.

2.2 Data Preparation

Prior to speech chunking, audio signals containing only the prepared monologues in the L1 and the L2 were automatically extracted and stored as separate audio files, to reduce lag during subsequent steps and to normalize amplitude by language[1]. The chunking of the spontaneous speech data was achieved by obtaining uninterrupted speech segments between 1.4 s and 1.6 s following Tilsen, Arvaniti (2013). This method reduces any variation that may be caused by differences in speech tempo and resolves the issue of time normalization, since chunk durations are uniformly distributed around 1.5 seconds with a ± 100 ms variation from this value (Tilsen, Arvaniti, 2013: 629).

The resulting speech signals were then prepared in *Praat* (Boersma, Weenink, 2021), following the initial stages of He, Dellwo's (2017) methodology. First, the DC bias was removed by subtracting the mean amplitude from the signal; then a higher-sampled amplitude envelope was created by low-pass filtering the full-wave rectified speech signal at 10 Hz [Hann filter, roll-off = 6 dB/octave]. Next, an intensity object was created in *Praat* (using the command To Intensity..., with Minimum pitch = 100 Hz; Time step = 0.0 s; Subtract mean = True). This command squares and windows the signal before creating the intensity object (Kaiser-Bessel window: $\beta = 20$; side lobe attenuation ≅ –190 dB; length: 32 ms). This series of signal manipulations results in the amplitude envelope of each signal and its intensity object containing intensity point values in time.

Next, the values in the intensity curve were linearly normalized within the range [0.01, 1] using the formula:

$$(1) \qquad I'(f) = (1 - 0.01)/(\max - \min)\times [I(f) - \min] + 0.01$$

Where $I'(f)$ and $I(f)$ refer to the normalized and original intensity value at frame index f; max and min refer to the maximum and minimum values of the original intensity curve, and 1 and 0.01 are the new maximum and minimum values of $I'(f)$. This procedure is analogous to the one employed by He et al. (2019) for the normalization of the first formant (F1). The authors proposed that the normalized curve maintains only information related to the trajectory of the (F1) curve that can be associated to speaker-specific articulatory gestures (He et al., 2019: 210).

Finally, the detection of intensity peaks and valleys was done semi-automatically. Instead of placing these points between pre-established syllable boundaries, an algorithm was created to automatically detect potential peak and valley points by surveying the amplitude envelope. After iterating through all points in the envelope collecting their intensity values in time, the algorithm determines if a point is a syllabic peak or valley by comparing successive intensity values. If the next value is larger than the current, the previous point is stored as a valley. Similarly, if the successive value is smaller than the previous point is stored as a peak. Next, all prospective pairs of peak and valley points are checked against each other. If their

[1] Normalizing the L1 and L2 data separately ensures that, for each language sample, no cross-linguistic influence would affect the analyses of the extracted measures of intensity dynamics.

difference is larger than a predetermined threshold (min 5 dB), they are considered as valid syllabic peaks or valleys. The output of this process is a series of intensity values in time of peak and valley points of each syllable in a continuous stretch of speech. These values were used in the automatic placement of peak and valley demarcation points in the intensity contour. These points were then manually checked to ensure correct placement.

2.3 Data Extraction

The intensity values of peaks and valleys were obtained at each of the demarcation points from the intensity curve using cubic interpolation, offering true continuity between the motion trajectories that pass through each peak and valley point. Next, positive dynamics ($\nu_{\mathrm{I}}[+]$) were computed by calculating the rate of intensity increased from a valley to its succeeding peak as follows:

(2) $$\nu_I[+] \stackrel{\text{def}}{=} \frac{(I_{\mathrm{P}} - I_{\mathrm{V}})}{(t_{\mathrm{P}} - t_{\mathrm{V}})}$$

Where I_{P} and I_{V} refer to the intensity values of the peak t_{P} and valley t_{V} points. Similarly, negative dynamics ($\nu_{\mathrm{I}}[-]$) were measured by calculating the rate of intensity decrease from a peak to its right-adjacent valley as shown in (3):

(3) $$\nu_I[-] \stackrel{\text{def}}{=} \frac{|I_{\mathrm{V}} - I_{\mathrm{P}}|}{(t_{\mathrm{V}} - t_{\mathrm{P}})}$$

Here the intensity values taken are absolute, since only the magnitude of the signal is of interest for the analyses (He, Dellwo, 2017: 490).

The distributions of positive and negative dynamics in a chunk of spontaneous speech were obtained by calculating the mean, standard deviation and Pairwise Variability Index, or PVI (Grabe, Low, 2002), of both types of dynamics from speakers' positive and negative slopes by language (min = 167, max = 586). The mean and standard deviation of each dynamic type display the central tendency and the overall dispersion of a speaker's intensity dynamics, respectively. The PVI conveys the amount of variability between successive syllables by computing and averaging the difference in duration between sequential intervals in an utterance (Grabe, Low, 2002). Following He, Dellwo's (2017) notation of these measures, MEAN_$\nu_{\mathrm{I}}[-]$, STDEV_$\nu_{\mathrm{I}}[-]$, and PVI_$\nu_{\mathrm{I}}[-]$ refer to negative dynamics and MEAN_$\nu_{\mathrm{I}}[+]$, STDEV_$\nu_{\mathrm{I}}[+]$, and PVI_$\nu_{\mathrm{I}}[+]$ to positive dynamics. These measures were stored per chunk per speaker in separate data subsets corresponding to language spoken; namely, English (EN) and Dutch (NL).

2.3 Statistical Analyses

According to He, Dellwo (2017) positive and negative measures may encode different types of information, which could be established if they are separated into two independent factors (He, Dellwo, 2017: 491). Therefore, factor analysis (FA) was employed on all measures of intensity dynamics in both language subsets to test whether positive and negative dynamics formed two independent categories (extraction method = principal components, eigenvalues ≧ 1; rotation method = Varimax with Kaiser normalization). Following, a multinomial logistic regression (MLR) was employed to assess how much inter-speaker variability is explained by the positive and negative dynamics. In this regression analysis, the measures of intensity dynamics were set as the numeric predictor variables and speaker as the nominal response variable.

Next, linear discriminant analysis (LDA) was employed to assess how well speakers can be discriminated based on positive and negative measures of intensity dynamics. The two types of dynamics were used as predictors in two separate analyses, one containing only positive measures and another containing only negative ones. In these analyses positive and negative measures are used as predictors, and speaker as the grouping variable (range = 1 to 51; within-group correlations with Fisher's coefficient; prior probability from group sizes; leave-one-out cross-validation). Post-hoc analyses were carried out with each individual measure of both types of dynamics to assess which of these measures were better predictors of class.

Subsequently, the effect of language on speakers' intensity dynamics was measured employing linear mixed-effects models (LME). In each model, built in a forward stepwise approach, one measure of intensity dynamics was assigned as the response variable, language as the main binary dummy fixed factor (0 = Dutch, 1 = English), and speaker as the random factor with by-language slopes (Bonferroni correction a posteriori). Assigning speakers as a random factor in the model increases the likelihood that a possible effect of language in the variation of these measures is genuine and can be generalized to all individuals fitting the sampling population (Walker, 2013: 454). These models were fitted in *R* (R Core Team, 2020) using the *nlme* package (Pinheiro, Bates, DebRoy & Sarkar, 2020) and using restricted maximum likelihood. Model selection was based on the best statistically significant model; in case of no significance in model fitness for a particular variable, the one with the lowest Log-likelihood value was selected.

3. *Results*

Table 1 provides a statistical description of the measures of intensity dynamics in both language subsets (N_{NL} = 1,843 and N_{EN} = 1,633). The U-test (Wilcoxon rank-sum test) for central tendency suggests that the center of each measure of intensity dynamics differs significantly between both languages. At first glance, STDEV_ν_I[-] shows relatively low dispersion in both language subsets compared to the other measures. Conversely, PVI_ν_I[−] shows the highest overall dispersion in L1 Dutch.

Although means and standard deviations are given for the sake of completeness, a better overview of the central tendency and dispersion of each measure of positive and negative dynamics can be seen in the values of medians and interquartile ranges (IQRs), since there are valid outliers present in the data which do affect the two former statistical descriptors.

Table 1 - *Means, standard deviations (SD), medians and IQRs for each measure of intensity dynamics by language. W indicates the result for the U-test for each measure of intensity dynamics. Significance values as follows:* ** $p < 0.01$; *** $p < 0.001$

	L1 Dutch			**L2 English**			
	Mean (SD)	**Median**	**IQR**	**Mean (SD)**	**Median**	**IQR**	**W**
*MEAN*_v_I[+]	4.25 (1.22)	4.19	1.59	4.12 (1.30)	4.03	1.76	1596146**
*STDEV*_v_I[+]	2.18 (.85)	2.13	1.15	2.11 (.94)	2.01	1.24	1598605**
*PVI*_v_I[+]	6.33 (1.67)	6.21	1.97	5.65 (1.42)	5.61	1.80	1850840***
*MEAN*_v_I[−]	3.23 (.89)	3.18	1.23	3.05 (.90)	3.00	1.27	1680520***
*STDEV*_v_I[−]	1.61 (.68)	1.55	.96	1.54 (.69)	1.48	.94	1595413**
*PVI*_v_I[−]	6.27 (1.73)	6.08	2.02	5.55 (1.39)	5.56	1.80	1867871***

3.1 Factor Analysis

For the Dutch subset the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (KMO = .513 > .5) and the Barlett's sphericity test (χ^2 = 4666.932, p < .0005) indicated that the data was suitable for factor analysis. The same was true for the English subset (KMO = .533 > .5; χ^2 = 3422.131, p < .0005).

The results on Table 2 show that two factors were extracted for the L2 English subset. Factor 1 includes all measures of negative intensity dynamics while Factor 2 includes all measures of positive dynamics. This outcome suggests that there is orthogonality between the measures of positive and negative dynamics, since they were classified into different factors. For the L1 subset these results were partially similar. While all measures of negative dynamics were classified into one factor, an additional measure of positive dynamics, PVI_v_I[+], was classified into the same factor (Factor 1). The remaining two positive measures were classified into a different factor (Factor 2). The reason why PVI_v_I[+] was classified alongside the measures of negative dynamics for this subset may lie in the very strong positive correlation (r = .80) between both positive and negative PVI measures.

Table 2 - *Factor loadings matrix after Varimax rotation. Absolute values under the threshold (.40) were suppressed from the table. The shaded cells indicate the highest loading values that classify a measure into a particular factor*

	L1 Dutch Factor Loadings		**L2 English Factor Loadings**	
	Factor 1	**Factor 2**	**Factor 1**	**Factor2**
*MEAN*_v_I[−]	.79		.83	
*STDEV*_v_I[−]	.79	−.72	.86	−.49
*PVI*_v_I[−]	.70		.57	
MEAN_v_I[+]		.82		.81
*STDEV*_v_I[+]		.90		.88
*PVI*_v_I[+]	.66		.43	.59
Eigenvalue	2.54	1.35	2.51	1.37
% of variance	42.30	22.57	41.86	22.88

3.2 Multinomial Logistic Regression

The results of the regression analysis (Table 3) show how much inter-speaker variability was explained by each measure of intensity dynamics. For the models concerning the L1 subset, 48% of between-speaker variability was explained by the combined positive measures and 52% by the combined negative measures. Similarly, in the L2 subset 49% of this variability was explained by the combined positive measures, and 51% by the combined negative measures. Among the negative measures, STDEV_v_I[−] seems to explain most variability in both subsets (Dutch = 28.95%, English = 18.35%); among the positive measures PVI_v_I[+] seems to better explain this variability in the L2 (18.30%), and STDEV_v_I[+] in the L1 (25.95%). Regarding both types of dynamics, STDEV_v_I[−] shows the greatest amount of variability between-speakers in the L1 model (28.95%), and in the L2 model (18.35%).

Table 3 - *Results of Multinomial Logistic Regression for both language subsets. −2LL provides model fit and* $\chi^2_{[df]}$ *tests how each measure explains the variance from the baseline model. Significance of reduced models:* $^*p < 0.0005$

	L1 Dutch			**L2 English**		
	−2LL	$\chi^2_{[df]}$	**Variability explained**	**−2LL**	$\chi^2_{[df]}$	**Variability explained**
(i) Model fitting information						
Null model	14346.47			12727.06		
Full model	13315.98	$1030.49_{[300]}$		11799.39	$927.67_{[300]}$	
(ii) Likelihood ratio test of each measure of intensity dynamics						
*MEAN*_v_I[−]	14016.91	$700.94_{[250]}^{*}$	22.70%	12428.18	$628.79_{[250]}^{*}$	14.58%
*STDEV*_v_I[−]	14209.84	$893.87_{[250]}^{*}$	28.95%	12590.72	$791.33_{[250]}^{*}$	18.35%
*PVI*_v_I[−]				12579.32	$779.93_{[250]}^{*}$	18.08%

	L1 Dutch			L2 English		
	−2LL	$\chi^2_{[df]}$	Variability explained	−2LL	$\chi^2_{[df]}$	Variability explained
MEAN_*v_I[+]*	14007.43	$691.45_{[250]}^{*}$	22.39%	12398.29	$598.90_{[250]}^{*}$	13.89%
STDEV_*v_I[+]*	14117.32	$801.34_{[250]}^{*}$	25.95%	12524.24	$724.85_{[250]}^{*}$	16.81%
PVI_*v_I[+]*				12588.50	$789.11_{[250]}^{*}$	18.30%
Σ		3087.60	100%		4312.91	100%

3.3 Linear Discriminant Analysis

Multivariate assumptions for data quality were met and the relatively large sample size for L1 Dutch (N = 1,843) and L2 English (N = 1,633) were deemed sufficient, suggesting that the analysis would be robust to some variations in data quality between groups and predictor variables, despite inequality in group sample sizes.

For the L2 subset all measures of intensity dynamics were entered in the LDA. For the L1 subset analysis the variables PVI_v_I[+] and PVI_v_I[−] were left out due to their very strong correlation (r = .80) and because when entered separately they did not affect the classification rates. To evaluate whether negative dynamics were a better predictor of speaker than positive dynamics, two separate analyses were carried out: one with the positive measures only (LDA[+]) and one only with the negative ones (LDA[−]). For both analyses all predictors were included simultaneously, since a stepwise approach did not improve classification rates. Prior odds were calculated from within-group sample sizes, and cross-validation was used.

The results from the analyses with combined measures showed that the overall percentage of correct classifications for Dutch was low, with virtually no difference in the classification performance involving positive or negative dynamics. For both analyses, LDA[+] and LDA[−], correct classification rates were ca. 4.8% (chance level = 1.9%). For English, LDA[−] had a higher rate of correct classification (4.8%) than LDA[+] (3.2%). Given this low classification performance, post-hoc analyses were carried out with each individual measure of positive and negative dynamics.

Table 4 - *LDA classification results (cross-validated) of individual measures per language subset. The column "Measure" displays which measure of intensity dynamics was used in the LDA. Chance level = 1.9%*

	L1 Dutch	L2 English
Measure		
MEAN_*v_I[+]*	4.8%	3.5%
STDEV_*v_I[+]*	3.4%	2.9%
PVI_*v_I[+]*	4.2%	3.5%
MEAN_*v_I[−]*	4.8%	4.4%
STDEV_*v_I[−]*	4.7%	4.3%
PVI_*v_I[−]*	4.0%	4.0%

The results of the post-hoc analyses (Table 4) did not show much improvement in the classification of speakers. However, this was not surprising, since having fewer predictors in the analysis would not necessarily improve classification. No discriminant function was able to classify speakers with more than 4.8% of accuracy. Nonetheless, these results made apparent that measures of central tendency of positive and negative dynamics (MEAN_ν_I[+] and MEAN_ν_I[−]) were better classifiers for the L1 subset (classification accuracy = 4.8%). For the L2 subset, STDEV_ν_I[−] (4.3%) and MEAN_ν_I[−] (4.4%) were better classifiers of group membership.

Table 5 - *Linear mixed-effects model to determine the effect of language on measures of intensity dynamics. Significance values as follows:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.
[a] *Significance values were corrected (Bonferroni)*

	MEAN_ν_I[−]		**STDEV_ν_I[−]**		**PVI_ν_I[−]**	
Estimation Method	REML		REML		REML	
Fixed–Effect Parameter	*Est. (SE)*	*95% CI*	*Est. (SE)*	*95% CI*	*Est. (SE)*	*95% CI*
β_0 *(Intercept)*	3.23 (.05)	[3.14, 3.32]	1.61 (.03)	[1.55, 1.66]	6.28 (.06)	[6.16, 6.40]
β_1 *(language)*[a]	− .17 (.03)***	[−.23, −.12]	− .07 (.02)*	[−.11, −.02]	− .72 (.05)***	[−.83, −.62]
Covariance Parameter	*Est.*	*95% CI*	*Est.*	*95% CI*	*Est.*	*95% CI*
σ^2	.85	[.83, .87]	.67	[.65, .68]	1.52	[1.48, 1.56]
σ^2_{int}	.30	[.24, .37]	.16	[.13, .21]	.34	[.27, .44]
	MEAN_ν_I[+]		**STDEV_ν_I[+]**		**PVI_ν_I[+]**	
Estimation Method	REML		REML		REML	
Fixed–Effect Parameter	*Est. (SE)*	*95% CI*	*Est. (SE)*	*95% CI*	*Est. (SE)*	*95% CI*
β_0 *(Intercept)*	4.27 (.07)	[4.13, 4.41]	2.19 (.04)	[2.11, 2.28]	6.34 (.06)	[6.23, 6.45]
β_1 *(language)*[a]	− .13 (.04)***	[−.21, −.05]	− .08 (.03)**	[−.14, −.02]	− .68 (.05)***	[−.78, −.58]
Covariance Parameter	*Est.*	*95% CI*	*Est.*	*95% CI*	*Est.*	*95% CI*
σ^2	1.17	[1.14, 1.20]	.86	[.84, .88]	1.52	[1.49, 1.56]
σ^2_{int}	.47	[.38, .59]	.26	[.20, .32]	.31	[.24, .41]

3.4 Linear Mixed-Effects Model

The results for the LME models used to explain the effect of language on each measure of intensity dynamics showed a strong main effect of language for all measures of intensity dynamics as seen on Table 5. Furthermore, each model's output for random effects (covariance parameter) suggests that the inclusion of the within-group predictor reduces the residual variability for PVI_v_{I}[−] and PVI_v_{I}[+] (σ^2 from 1.56 to 1.52). That is, for these measures setting language as a predictor explains that variability thought to be within-speaker was actually due to differences across languages. It is noteworthy to mention that the LME models were significantly improved in model fitting for all measures by adding random slopes. However, a strong main effect of language for all measures was only observed when the random effects part did not include the interaction between speaker and language. Therefore, here I only reported the models without the 2-way interaction.

4. *Discussion*

The goals of this study were to investigate (i) whether between-speaker variability in intensity dynamics would be present in spontaneous production of L1 Dutch – L2 English speakers and (ii) whether language spoken would influence intensity dynamics. The results reported here indicated that inter-speaker variability was indeed reflected in the measures of intensity dynamics, despite the language spoken by the individuals, contributing to the claim that differences in the production of this feature could also be attributed to speaker-specific biomechanical characteristics (He, Dellwo, 2017).

Interestingly, both MLR and FA analyses did not fully replicate the results found for L1 Zurich German speakers, for which (a) the distribution of the variability in positive and negative dynamics was highly unbalanced, and (b) both types of dynamics were perfectly classified into different factors. Regarding the MLR, one possible explanation for the results in this study relates to the type of data used. Unlike read speech, spontaneous utterances are believed to display larger variation of articulatory patterns and different speech rates (De Nil, Abbs, 1991; Illa, Ghosh, 2020), both of which would affect intensity dynamics. Therefore, a more balanced distribution of between-speaker variability in positive and negative dynamics in both languages could be attributed to inherent differences between spontaneous and elicited speech (DiCanio, Nam, Amith, García & Whalen, 2015; Simpson, 2013). Nonetheless, the results presented here did follow in both languages the earlier reported tendency that negative dynamics could explain more variability between speakers than its positive counterpart.

As for the results of the FA, differences in speaking style could also explain why some measures of positive dynamics were classified with negative dynamics for the L1. This seems to indicate that, at least for Dutch, the information encoded by both types of dynamics is not completely orthogonal in spontaneous speech since the results showed no evidence that PVI of positive dynamics would be under more prosodic control. As for the L2, this assumption cannot be extended, since both types of dynamics followed the tendency reported by He, Dellwo (2017), where positive dynamics may have been under more prosodic control than negative dynamics. Another interpretation of the different results regarding the L1 and the L2 could be related to the notion that non-native speech may be more carefully produced than native speech, and therefore leave less room for variation in positive dynamics.

Although the results did not completely follow earlier trends, both types of intensity dynamics still significantly explained inter-speaker variability. Taking this into account, the results of the LDA provided a practical picture of the usability of this measure by assessing the discriminative power of intensity dynamics in spontaneous speech data. The power of each measure did not depend on whether the speaker used their L1 or L2. Overall, the higher classification rates for negative dynamics in both languages reflect the results in the MLR: negative dynamics explained more inter-speaker variability than positive dynamics. Interestingly, this was more strongly the case for the L1 than for the L2. This behavior in the L2 seems again to suggest that a more balanced amount of variability explained by positive and negative dynamics could be linked to careful productions of the L2 (Kormos, 2006).

Although overall both types of dynamics seemed to be poor discriminators for spontaneous speech data, a careful inspection of the results made evident that for some speakers negative dynamics were a better predictor in the L1 than in the L2. These results seem to indicate that language may affect measures of positive and negative intensity dynamics differently in each speaker. This was visible in the confusion matrices, where some speakers had higher correct classification scores in the L2 than in the L1. As to why this was observed remains an open question for now. Although generally, the inspection of the confusion matrices strengthens the assumption that language may affect the discriminative power of these measures differently across speakers.

The results of the LME models confirmed the assumption that language would somewhat influence intensity dynamics. These results showed a significant effect of language on the temporal organization of intensity contours and were interpreted as systematic differences between the rhythmic characteristics of Dutch and English. Since both languages belong to the West Germanic language family, they share many similarities. However, there are also considerable differences between these languages in terms of the phonological and phonetic parameters employed during speech production (Hirst, Di Cristo, 1998; Alber, 2020; Page, 2020), which characterizes the acoustic parameters of each language.

Finally, the fact that intensity dynamics still displayed enough between-speaker variability in both languages strengthens the claim that speaker-specificity may not be constrained to the L1 (Bradlow, Blasingame & Lee, 2018; Vaughn, Baese-Berk & Idemaru, 2019). Moreover, since it has been proposed that the L1 phonological inventory of a speaker may influence their L2 production (Flege, 1995; Best, Tyler, 2006; Escudero, 2009), the results presented here also seem to indicate that not only static information is carried over from the L1 to the L2, but also dynamic information, which has also been proposed to be stored in the phonological system of a speaker (Schwartz, Kaźmierski, 2019).

5. *Limitations and future research*

Although the results of this study provide a significant understanding of speaker-specific influences on the spontaneous production of intensity dynamics in a native and non-native language, the limitations need to be addressed. First, adapting He, Dellwo's (2017) method for the extraction of intensity dynamics could be lacking in the sense that it was intended for prepared rather than spontaneous speech. Moreover, the chunking method, although previously employed in the investigation of rhythmic characteristics in spontaneous speech, could have influenced the correlation of some measures, since longer continuous stretches of speech were chunked to reduce variation in sample length.

Secondly, to reliably evaluate whether differences between spontaneous and read speech are significant for intensity dynamics in L1 Dutch, research with different speech styles needs to be conducted. Likewise, this should be considered for the L2. In addition, while the hypotheses were confirmed for proficient L2 speakers, it is also wise to test whether the results would be similar for beginner and intermediate L2 speakers. It is assumed that the L1 will strongly influence L2 productions of these speakers; however, one should assess whether the level of control over the articulatory gestures governing intensity dynamics is indeed correlated to the degree of L2 knowledge.

Moreover, I should emphasize the need to cross-validate the obtained results. Kraayeveld (1997: 120) pointed out that time-integrated acoustic measures not only depend on the speaker, but they also vary over time. Consequently, cross-validation is necessary to assess whether the speaker-specific information present in intensity dynamics would indeed remain consistent for a speaker over time. Ultimately, after considering the study's limitations, it remains evident that speaker-specific characteristics in intensity dynamics were found in both languages. Therefore, future research should seek to investigate to what degree intensity dynamics in one language would allow the identification of a speaker in another.

Acknowledgements

This paper is based on the author's MA thesis supervised by Dr. Willemijn Heeren at Leiden University in the Netherlands.

Bibliography

Alber, B. (2020). Word Stress in Germanic. In Putnam, M.T., Page, B.R. (Eds.), *The Cambridge Handbook of Germanic Linguistics*. Cambridge: Cambridge University Press, 73-96. https://doi.org/10.1017/9781108378291.005

Best, C.T., Tyler, M.D. (2006). Nonnative and second-language speech perception: Commonalities and complementarities. In Munro, M.J., Bohn, O.S. (Eds.), *Second* *language speech learning: The role of language experience in speech perception and production*. Amsterdam: John Benjamins, 13-34. https://doi.org/10.1075/lllt.17.07bes

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer. [Computer program] Version 6.1.41, retrieved April, 12 2021 from http://www.praat.org/.

Bradlow, A.R., Blasingame, M. & Lee, K. (2018). Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech. In *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 9(1), 1-20. https://doi.org/10.5334/labphon.137

Chandrasekaran, C., Friston, K.J., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A.A. (2009). The Natural Statistics of Audiovisual Speech. In *PLoS Computational Biology*, 5(7), 1-18. https://doi.org/10.1371/journal.pcbi.1000436

Coulthard, M., Johnson, A. & Wright, D. (2016). *An introduction to forensic linguistics: Language in evidence*. London: Routledge. https://doi.org/10.4324/9781315630311

De Nil, L.F., Abbs, J.H. (1991). Influence of speaking rate on the upper lip, lower lip, and jaw peak velocity sequencing during bilabial closing movements. In *The Journal of the Acoustical Society of America*, 89(2), 845-849. https://doi.org/10.1121/1.1894645

DiCanio, C., Nam, H., Amith, J.D., García, R.C. & Whalen, D.H. (2015). Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec. In *Journal of Phonetics*, 48, 45-59. https://doi.org/10.1016/j.wocn.2014.10.003

Escudero, P. (2009). The linguistic perception of "similar" L2 sounds. In Boersma, P., Hamann, S. (Eds.), *Phonology in perception*. Berlin, New York: Mouton de Gruyter, 151-190. https://doi.org/10.1515/9783110219234.151

Flege, J.E. (1995). Second-language speech learning: Theory, findings, and problems. In Strange, W. (Ed.), *Speech perception and linguistic experience: Issues in Cross-Language Research*. Timonium: York Press, 233-277.

Fry, D.B. (1979). *The physics of speech*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139165747

Gold, E., French, P. (2011). International practices in forensic speaker comparison. In *The International Journal of Speech, language and the law,* 18(2), 293-307. https://doi.org/10.1558/ijsll.v18i2.293

Goldstein, U.G. (1976). Speaker-identifying features based on formant tracks. In *The Journal of the Acoustical Society of America*, 59(1), 176-182. https://doi.org/10.1121/1.380837

Grabe, E., Low, E.L. (2002). Durational variability in speech and the rhythm class hypothesis. In Warner, N., Gussenhoven, C. (Eds.), *Papers in laboratory phonology* 7. Berlin: Mouton de Gruyter, 515-546. https://doi.org/10.1515/9783110197105.2.515

He, L., Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. In *The Journal of the Acoustical Society of America,* 141(5), 488-494. https://doi.org/10.1121/1.4983398

He, L., Zhang, Y. & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. In *The Journal of the Acoustical Society of America*, 145(3), 209-214. https://doi.org/10.1121/1.5093450

Hirst, D., Di Cristo, A. (1998). A survey of intonation systems. In Hirst, D., Di Cristo, A (Eds.), *Intonation systems: a survey of twenty languages*. Cambridge: Cambridge University Press, 56-77.

Hollien, H. (1990). *The acoustics of crime: The new science of forensic phonetics*. New York: Springer Science & Business Media. https://doi.org/10.1007/978-1-4899-0673-1

Hixon, T.J., Weismer, G. & Hoit, J.D. (2020). *Preclinical speech science: Anatomy, physiology, acoustics, and perception*. San Diego: Plural Publishing.

Illa, A., Ghosh, P.K. (2020). The impact of speaking rate on acoustic-to-articulatory inversion. In *Computer Speech & Language*, 59, 75-90. https://doi.org/10.1016/j.csl.2019.05.004

Kitamura, T., Akagi, M. (2007). Speaker individualities in speech spectral envelopes and fundamental frequency contours. In Müller, C. (Ed.), *Speaker Classification II*. Springer, Berlin, Heidelberg, 157-176. https://doi.org/10.1007/978-3-540-74122-0_14

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah: Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203763964

Kraayeveld, H. (1997). *Idiosyncrasy in prosody: speaker and speaker group identification in Dutch using melodic and temporal information*. PhD dissertation, University of Nijmegen.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Marchal, A. (2009). *From speech physiology to linguistic phonetics*. London: John Wiley & Sons.

McDougall, K. (2007). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. In *International Journal of Speech Language and the Law*, 13(1), 89-126. https://doi.org/10.1558/ijsll.v13i1.89

Orr, R., Quené, H. (2017). D-LUCEA: Curation of the UCU Accent Project Data. In Odijk, J., van Hessen, A. (Eds.), *CLARIN in the Low Countries*. London: Ubiquity Press, 181-193.

Orr, R., Quené, H., Beek, R.V., Diefenbach, T., Leeuwen, D.A.V. & Huijbregts, M. (2011). An international English speech corpus for longitudinal study of accent development. In *Proceedings of INTERSPEECH-2011*, Florence, Italy, 27-31 August 2011, 1889-1892.

Page, B.R. (2020). Quantity in Germanic Languages. In Putnam, M.T., Page, B.R. (Eds.), *The Cambridge Handbook of Germanic Linguistics*. Cambridge: Cambridge University Press, 97-118. https://doi.org/10.1017/9781108378291.006

Perrier, P., Winkler, R. (2015). Biomechanics of the orofacial motor system: Influence of speaker-specific characteristics on speech production. In Fuchs, S., Pape, D.P. & Perrier, P. (Eds.), *Individual Differences in Speech Production and Perception*. Frankfurt: Peter Lang, 223-254.

Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. (2020). nlme: Linear and Nonlinear Mixed Effects Models. [R package] Version 3.1-148, retrieved May, 15 2020 from https://CRAN.R-project.org/package=nlme.

R Core Team (2020). R: A language and environment for statistical computing. [Computer program] Version 4.0.0 retrieved April, 24 2020 from https://www.R-project.org/.

Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor and Francis.

Schwartz, G., Kaźmierski, K. (2019). Vowel dynamics in the acquisition of L2 English– an acoustic study of L1 Polish learners. In *Language Acquisition*, 27(3), 227-254. https://doi.org/10.1080/10489223.2019.1707204

Simpson, A.P. (2013). Spontaneous speech. In Jones, M., Knight, R.A. (Eds.), *Bloomsbury Companion to Phonetics*. London: Bloomsbury Publishing Plc, 155-169. http://dx.doi.org/10.5040/9781472541895.ch-010

Stevens, K.N. (2000). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. In *Philosophical Transactions of the Royal Society B: Biological Sciences*, 335(1273), 71-78.

Tilsen, S., Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. In *The Journal of the Acoustical Society of America*, 134(1), 628-639. https://doi.org/10.1121/1.4807565

Titze, I.R., Palaparthi, A. (2018). Radiation efficiency for long-range vocal communication in mammals and birds. In *The Journal of the Acoustical Society of America*, 143(5), 2813-2824. https://doi.org/10.1121/1.5034768

Vaughn, C., Baese-Berk, M. & Idemaru, K. (2019). Re-examining phonetic variability in native and non-native speech. In *Phonetica*, 76(5), 327-358. https://doi.org/10.1159/000487269

Walker, J.A. (2013). Variation analysis. In Podesva, R.J., Sharma, D. (Eds.), *Research methods in linguistics*. Cambridge: Cambridge University Press, 440-459. https://doi.org/10.1017/CBO9781139013734.023

Zuo, D., Mok, P.P.K. (2015). Formant dynamics of bilingual identical twins. In *Journal of Phonetics*, 52, 1-12. https://doi.org/10.1016/j.wocn.2015.03.003