BETWEEN-SPEAKER SYLLABLE INTENSITY VARIABILITY IN PERSIAN

Homa Asadi[1], Batool Alinezhad

Department of Linguistics
University of Isfahan, Iran
{h.asadi; b.alinezhad}@fgn.ui.ac.ir

ABSTRACT

This study examines speaker-specific temporal features with a particular focus on the variability of syllable intensity in a Persian-speaking population. Two types of intensity variability measures (mean and peak) between syllables were examined as a function of speaker in two Persian databases with different sources of within-speaker acoustic variability. Results showed that the intensity measures that take into account the difference between all intervals in an utterance were better at revealing between-speaker variability than the measures based on the difference between successive intervals. Peak intensity measures were also found to be more speaker-specific. We also found that the degree of between-speaker rhythmic variability in terms of syllable intensity was not affected by the language-specific features of Persian. However, the discriminatory power of the intensity measures was reduced in the database with extreme rate variability. We discuss how results could be applied in forensic speaker comparison.

Keywords: speaker discrimination, intensity measures, vocal variability.

I. INTRODUCTION

Acoustic characteristics of voice with high variability between speakers and low variability within speakers are crucial to forensic speaker comparison (FSC) [1, 2, 3, 4]. In FSC, two samples of voices, a known and an unknown (disputed) sample, are compared to estimate the probability that the same speaker has produced the speech samples (same-speaker hypothesis) versus the probability that the speech samples have come from two different speakers (different-speaker hypothesis) [3]. This goal can be achieved by learning about acoustic parameters that discriminate between speakers well and show high stability within each speaker.

Previous studies provided strong evidence that acoustic measures of speech rhythm based on the durational characteristics of consonantal and vocalic intervals as well as on the syllable intensity characteristics can help distinguish different speakers [5, 6, 7, 8]. It has been demonstrated that anatomical differences in the structural dimensions of speakers' articulators, as well as acquired idiosyncratic ways of operating their speech articulators to produce sounds, will significantly alter the acoustic characteristics of speech and could, therefore, result in a high degree of between-speaker variability in temporal aspects of speech [5, 6, 7, 9].

Rhythmic variability between speakers is largely influenced by intensity in acoustic signals. Empirical studies based on audiovisual experiments have shown that mouth-opening size and vocal intensity are highly correlated and that articulatory movements of the lips and mouth opening are reflected in the temporal structure of the amplitude envelope [10, 11, 12, 13].

Acoustically, between-speaker variability in syllable intensity was first investigated by [7, 14]. They argued that individual articulatory movements could lead to a high degree of speaker specificity, which can be reflected in the intensity aspects of speech signals. They applied different intensity measures to the speech corpora with speakers from Zürich German and North German and found enormous variability between speakers in the two databases.

In addition to anatomical and learned differences that affect temporal characteristics of speech, cross-linguistic differences may also play a role in syllable intensity variability. It has been demonstrated that a language with a more phonotactically complex structure typically has a higher degree of intensity variability than languages with simpler structures [15]. Additionally, languages that allow vowel reduction have higher variability of syllable intensity because reduced vowels have lower amplitude envelopes, resulting in lower intensity levels in terms of mean intensity or peak intensity [16]. Another important factor affecting intensity is related to how lexical stress is signaled in a language. Languages that use intensity as an acoustic cue for signaling lexical stress have higher levels of intensity variability than those that do not utilize intensity for the production of stress [17].

Persian has a simple syllable structure of CV(C)(C), and it does not have a vowel reduction

pattern [18, 19, 20, 21, 22, 23]. In terms of stress pattern, duration is the most reliable correlate of stress in Persian, while intensity is a poor marker of stress position [24]. We thus hypothesize that the degree of between-speaker rhythmic variability in terms of syllable intensity is prone to be decreased in Persian because of less complex syllable structure, little to no vowel reduction, and different phonetic realization of stress.

In line with [7] and [14], we have selected two types of intensity measures and analyzed between- and within-speaker variability, posing the following questions:

1) Do intensity measures vary between speakers in Persian?

2) Do intensity measures remain stable across different sources of within-speaker variability?

3) Which intensity measures explain possible between-speaker variability best?

2. METHOD

2.1. Participants and task

Two speech corpora with different sources of within-speaker acoustic variability were collected for this study. In the first corpus (hereafter non-contemporaneous corpus), 20 male native speakers of Persian (age mean=31.2, SD = 4.3) were recorded in two sessions separated by a two-month time-lapse. Speakers were equipped with a fixed microphone and asked to read 40 Persian sentences at a normal rate with a three-second pause between sentences. Recording sessions were conducted in a soundproof booth with a sampling rate of 44.1 kHz and a quantization of 16 bits. For the second corpus (hereafter tempo corpus), following the procedure used in the collection of the BonnTempo corpus in German [25], 10 male native speakers of Persian (age mean=34.3, SD = 3.6) were instructed to read *The North Wind and the Sun* in Persian at five different speaking speeds (normal, slow, slower, fast, and fastest possible). Before each recording session, participants were asked to read the text several times to familiarize themselves with the passage. Speakers were then asked to slow their pace in two steps and then to read the text faster and as fast as possible. This resulted in strong syllable rate variability across the five different reading passages. The recording location and setup were the same as in the non-contemporaneous corpus.

2.2 Acoustic parameters

Speech materials were acoustically analyzed using Praat [26]. Speech tokens were annotated in segments, syllables, and peak tiers. We developed two sets of intensity variability metrics following the procedure used in [7, 14]. From the syllable tier, we calculated the stdevM, varcoM, nPVIm and rPVIm, while stdevP, varcoP, nPVIp and rPVIp were calculated from the peak syllable tier. To quantify the mean syllable intensity, we divided the sum of the intensity values between the onset and offset of a syllable by its duration. Peak intensity was calculated at the syllable peak point interpolated with the cubic function. We calculated the global quantification of intensity measurements based on the standard deviations of the mean and peak intensity of the syllables in an utterance. Local intensity variations were calculated by measuring syllable-to-syllable intensity differences. A detailed description of the intensity measurements is given in the appendix.

2.3. Statistical analyses

All statistical analyses were performed using R (R core Team, 2022) version 4.2.2 [27]. First, we conducted a principal component analysis (PCA) to see whether the different intensity measures formed independent categories (eigenvalues $\geqq 1$; rotation method = varimax). Second, we constructed a multinomial logistic regression (MLR) model on the collected speech data to address the question of which intensity measures could better explain the variability between speakers. We modeled the acoustic parameters as predictor variables and the speaker as a nominal response variable. The proportion of between-speaker variability explained by intensity measures was calculated using the likelihood ratio χ^2 of each acoustic parameter divided by the sum of the likelihood ratio χ^2s of all parameters. Furthermore, linear mixed-effects models were run to analyze the significance of within-speaker acoustic variability, i.e., repetition and tempo on intensity measures in our datasets. In the non-contemporaneous corpus, repetition was entered into the model as a fixed effect, and the speaker and sentence were treated as random factors, whereas in the tempo corpus, the tempo was considered as a fixed effect and the speaker and sentence as random factors.

3. RESULTS

3.1. Principal component analysis (PCA)

The PCA results from Table 1 show that three components were extracted for the non-contemporaneous corpus, with component 1 including all measures of mean intensity, while components 2 and 3 were primarily based on peak intensity measures. For the tempo corpus, two components were extracted. The first component contains measures computed over the entire

utterance, while local measures have emerged in the second component. This suggests that the different types of intensity measurements contain complementary information.

Table 1: PCA table showing a correlation matrix for variables loaded on the PCs extracted from the data analysis.

	Non-contemporaneous corpus		
	Comp1	Comp2	Comp3
rPVIm	0.952		
nPVIm	0.950		
varoM	0.786		
stdevM	0.772		
stdevP		0.921	
varcoP		0.920	
nPVIp			0.973
rPVIp			0.965
variance	**0.39**	**0.29**	**0.25**
	Tempo corpus		
	Comp1	Comp2	
varcoP	0.952		
stdevP	0.950		
stdevM	0.786		
varcoM	0.772		
nPVIm		0.892	
rPVIm		0.884	
nPVIp		0.885	
rPVIp		0.846	
variance	**0.43**	**0.38**	

3.2. Multinomial logistic regression

MLR analysis results show that the speaker's effect was significant in all investigated intensity measures. Nonetheless, these measures were not balanced in explaining between-speaker variability. As shown in Tables 2 and 3, the strongest effects were found for varcoP in both speech corpora. Based on the results, most acoustic variation is explained by peak and global measures.

Table 2: Summary of the results of MLR on intensity measures for non-contemporaneous corpus.

Intensity measures	-2 Log Likelihood of Reduced Model	$\chi 2$ [df]	P	Variability explained
stdevM	11476.703	193.510 [19]	<0.0001	12%
varcoM	11465.566	182.373 [19]	<0.0001	11.3%
rPVIm	11381.953	98.760 [19]	<0.0001	6.1%
nPVIm	11381.067	97.874 [19]	<0.0001	6.1%
stdevP	11638.987	355.795 [19]	<0.0001	22.1%
varcoP	11649.108	365.916 [19]	<0.0001	22.8%
rPVIp	11433.896	150.704 [19]	<0.0001	9.3%
nPVIp	11441.893	158.701 [19]	<0.0001	9.8%

Table 3: Summary of the results of MLR on intensity measures for tempo corpus

Intensity measures	-2 Log Likelihood of Reduced Model	$\chi 2$ [df]	P	Variability explained
stdevM	1187.803	130.399 [9]	<0.0001	13.9%
varcoM	1190.133	132.729 [9]	<0.0001	14.1%
rPVIm	1129.827	66.544 [9]	<0.0001	7.1%
nPVIm	1123.948	72.423 [9]	<0.0001	7.7%
stdevP	1298.924	241.52 [9]	<0.0001	25.8%
varcoP	1299.406	242.002 [9]	<0.0001	25.8%
rPVIp	1083.017	25.613 [9]	<0.0001	2.7%
nPVIp	1081.619	24.215 [9]	<0.0001	2.6%

3.2. Linear mixed-effects models

The results of within-speaker occasion-to-occasion variability analysis showed that the variability of the tested parameters as a function of repetition is not significant ($p>0.05$). Results also revealed that the effect of speaker on the intensity measures was significant across different speech rates. Further, we analyzed the effect of speaker at the five different speech rates separately. The result showed that speakers are differentiated well when speaking at normal, slow, fast and fastest rates, but they behave similarly when speaking at the slowest rate. Post hoc analyses using Bonferroni adjusted pairwise t-test were also applied to quantify the differences between speakers in both corpora. It has been shown that the number of significant paired comparisons is higher in the non-contemporaneous corpus. For example, for nPVIp, 215 of the 361 (59%) comparisons were significant ($p< 0.0001$) in the non-contemporaneous corpus, while 19 out of the 81 (23%) possible paired comparisons in the tempo corpus were significant ($p<0.05$). Comparing the results of the tempo corpus with those obtained in the non-contemporaneous corpus shows that between-speaker variability is drastically reduced in the dataset with high prosodic variability. Figure 1 shows the boxplot of between- and within-speaker variability for varcoP in both investigated corpora.

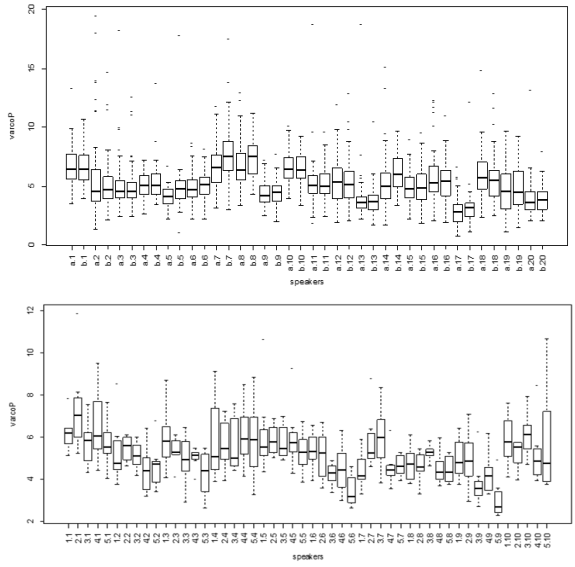

Figure 1: Boxplots of between- and within-speaker variability for varcoP in non-contemporaneous (top) and tempo corpus (bottom) (Rerecording sessions are shown with a and b. Speech rates are shown as follows: slow=1, slowest=2, normal=3, fast=4, fastest=5).

4. DISCUSSION AND CONCLUSION

In this study, we investigated between-syllable intensity variability as a source of between-speaker rhythmic variability in two Persian speech corpora with various sources of within-speaker variability. Our findings showed that intensity measures vary considerably and consistently among Persian speakers in both speech corpora. Results of our study based on the non-contemporaneous corpus replicate the previous studies on German and Swiss German databases [7]. As such, we hypothesized that language-dependent features, such as syllable structure complexity, vowel reductions, and the use of different acoustic cues to signal lexical stress, may affect speaker discrimination strength. However, our findings proved that such features do not affect the discriminatory power of intensity measures. This implies that intensity measures can discriminate between speakers regardless of their language of communication. We also found that speakers' performances were similar across their two recording sessions which is further indicative that intensity measures are robust to within-speaker acoustic variability caused by time-lapsing. Additionally, the results suggested that despite high prosodic variability in tempo corpus, speakers exhibited differences in intensity measures. Nevertheless, the obtained results are not totally in line with the previous studies [5, 7]. In Persian, intensity measures performed well in discriminating speakers when they were speaking at normal, fast, very fast and slow rates. Still, those measures were unable to identify speakers when they reduced their speed to the slowest possible rate. This is not the case for German speakers. Intensity measures could discriminate among German speakers in the BonnTempo corpus, a dataset of the same speech material with high within-speaker speech rate variability. A possible reason for the different performance of intensity measures at the slowest rate in German and Persian may be that Persian speakers uttered the passage with varying strategies of speed and loudness, which subsequently led to some changes in the intensity scores. On the one hand, speakers usually utilize less energy at a slow rate in moving their articulatory organs of speech, especially lips and the blade of the tongue. On the other hand, they have less control over the idiosyncratic muscular movement of speech organs. These will result in less speaker-specific information in the intensity variability across syllables in the utterances at the slowest rate.

In line with [7, 14], we also found that the measures that consider the difference between all the intervals in an utterance or sentence perform better in revealing between-speaker variability than those based on the difference between consecutive intervals. Our results also showed that the peak measures collectively contain more between-speaker variability than the mean measures in both databases. Intensity peaks are strongly correlated with articulatory correlates of jaw and tongue tip maximum displacements [13, 28, 29]. Thus, speaker-specific articulatory behavior may be more influential on the intensity peak that occurs when the mandible or tongue tip reaches its maximum displacement [14]. VarcoP yielded the best result for speaker discrimination in both Persian databases. This accords with the findings of [7], who also reported varcoP as a powerful speaker-specific variable with robustness against high prosodic variability. According to previous and current studies, we can conclude that varcoP is a powerful universal speaker-specific feature that goes beyond typological differences across languages.

Some implications of our results are relevant to the field of forensic phonetics. This study supports the notion that intensity measures can be used to determine speaker individuality. According to previous studies demonstrating that intensity measures can discriminate speakers in German and Swiss German, the results of this study in Persian suggest that speech rhythm intensity measures can be characterized as language-independent measures capable of being used when speaker-specific rhythms are not known.

6. APPENDIX

– The global intensity measures:
- stdevM: the standard deviation of average syllable intensity levels;
- stdevP: the standard deviation of syllable peak intensity levels;
- varcoM: the variation coefficient of average syllable intensity levels
- varcoP: the variation coefficient of syllable peak intensity levels

– The local intensity measures:
- rPVIm: the raw pairwise variability of adjacent mean syllable intensity levels;
- rPVIp: the raw pairwise variability of adjacent syllable peak intensity levels;
- nPVIm: the normalized pairwise variability of adjacent mean syllable intensity levels;
- nPVIp: the normalized pairwise variability of adjacent syllable peak intensity levels.

The formulae for calculating intensity variability can be found in [7, 14].

7. REFERENCES

[1] Wolf, J. 1972. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*. 51(6B), 2044–2056.

[2] Nolan, F. 1983. *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

[3] Rose, P. 2003. *Forensic Speaker Identification*. New York: Taylor & Francis.

[4] Morrison, G. S. 2010. Forensic Voice Comparison, In I. Freckelton & H. Selby, *Expert Evidence*. Sydney: Thomson Reuters.

[5] Dellwo, V., Leemann, A., Kolly, M. 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *Journal of the Acoustical Society of America*. 137, 1513–1528.

[6] Leemann, A., Kolly, M., Dellwo, V. 2014. Speaker-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Science International*, 238, 59–67.

[7] He, L., Dellwo, V. 2016. The role of syllable intensity in between-speaker rhythmic variability. *The International Journal of Speech, Language and the Law*. 23, 243–273.

[8] Asadi, H., Nourbakhsh, M., He, L., Pellegrino, E., Dellwo, V. 2018. Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persian reveals. *International Journal of Speech, Language and the Law*. 25(2), 151–174.

[9] Dellwo, V., Huckvale, M. and Ashby, M. 2007. How is individuality expressed in voice? An introduction to speech production and description for speaker classification". In C. Müller (Ed), *Speaker Identification 1*, 1-20, Berlin: Springer Verlag.

[10] Summerfield Q. 1992. Lipreading and audio-visual speech perception. *Philosophical transactions of the Royal Society of London Biological Sciences*. 335(1273):71–8.

[11] Grant KW, Seitz PF. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*. 108(3),1197–208.

[12] Garnier, M., Wolfe, J., Henrich, N., Smith, J. 2008. Interrelationship between vocal effort and vocal tract acoustics: a pilot study. *Proceedings of INTERSPEECH* Brisbane, 2302–2305.

[13] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A.A. 2009. The natural statistics of audiovisual speech. *PLoS Computational Biology*, 5(7), e1000436.

[14] He, L., Dellwo, V. 2014. Speaker idiosyncratic variability of intensity across syllables. *Proceedings of INTERSPEECH* Singapore.

[15] Prieto, P., del Mar Vanrell, M., Astruc, L., Payne, E., Post, B. 2012. Phonotactic and phrasal properties of speech rhythm: Evidence from Catalan, English and Spanish. *Speech Communication*. 54, 681–702.

[16] He, L. 2017. *Speaker Idiosyncratic Intensity Variability in the Speech Signal*. Ph.D. Dissertation, University of Zurich.

[17] Wang, Q. 2008. L2 stress perception: The reliance on different acoustic cues. *Speech Prosody* Campinas, 635–638.

[18] Windfuhr, G. L. 1979. *Persian grammar: History and state of its study*. New York: Mouton de Gruyter.

[19] Lazard, G. 1992. *Grammar of contemporary Persian*. Mazda Publishers.

[20] Sheikh Sangtajan, Sh., Bijankhan, M. 2010. The study of vowel reduction in Persian spontaneous speech. *Journal of Research in Linguistics*, 2(1): 35–48.

[21] Bijankhan, M. 2018. Phonology. In A. Sadeghi and P. Shabani-Jadidi (eds) *The Oxford Handbook of Persian Linguistics*; 111–141. Oxford: Oxford University Press.

[22] Yavaş, M. 2011. *Applied English phonology*, United Kingdom: Wiley-Blackwell.

[23] Sadeghi, V. 2015. A phonetic study of vowel reduction in Persian, *Language Related Research.* 30, 165–187.

[24] Sadeghi, V. 2011. Acoustic correlates of lexical stress in Persian. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)* Hong Kong 1738–1741.

[25] Dellwo, V. 2010. *Influences of speech rate on the acoustic correlates of speech rhythm: an experimental phonetic study based on acoustic and perceptual evidence*. Ph.D. Dissertation, Bonn University.

[26] Boersma, P., & Weenink, D. (2022) Praat: Doing Phonetics by Computer. Version 6.2.14, retrieved 24 March, 2021 from http://www.praat.org/.

[27] R Core Team R, A Language and Environment for Statistical Computing (version 4.2.2). 2022. R Foundation for Statistical Computing. http://www.Rproject.org.

[28] Birkholz, P., Kröger, B. J., Neuschaefer-Rube, C. 2011. Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*. 9(5), 1422–1433.

[29] Erickson, D., Kim, J., Kawahara, S., Wilson, I., Menezes, C., Suemitsu, A., Moore, J. 2015. Bridging articulation and perception: The C/D model and contrastive emphasis. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow.