Czech Voiced Labiodental Continuant Discrimination from Basic Acoustic Data

Radek Skarnitzl, Jan Volín

Institute of Phonetics
Charles University in Prague, Faculty of Philosophy & Arts
radek.skarnitzl@ff.cuni.cz

Abstract

The Czech voiced labiodental continuant has a special position within the system of Czech consonants. Due to its past it sometimes behaves phonologically as a sonorant, while its canonical articulatory qualities place it among obstruents. It is quite difficult to decide how much of its sound is sonorant and how much is obstruent by mere auditory analysis. We investigated whether basic acoustic parameters extracted from fluent reading material can help to discriminate if the Czech [v] leans more towards the obstruent or the sonorant class of consonants. Dynamic profile data were processed by factor analysis and, together with duration and harmonicity measurements, used as an input for cluster analysis, discriminant analysis, and classification-tree analysis. The results show that acoustically, the Czech intervocalic [v] in connected speech is more in line with its sonorant past rather than its obstruent presence.

1. Introduction

The consonantal inventory of modern Czech consists of two natural classes - obstruents and sonorants. The obstruents form pairs where specific place and manner of articulation is represented by one voiceless and one voiced segment. In certain positions these segments trigger or undergo assimilation of voicing. The sonorants, on the other hand, do not come in pairs, they are always voiced, and they neither trigger nor undergo assimilation of voicing. One exception to this neat arrangement exists, however. The labiodental pair /f, v/ is formed by a voiceless /f/ which behaves as a proper obstruent in that it both triggers and undergoes assimilation, and a voiced /v/ which only undergoes but does not trigger assimilation of voicing.

For explanation of this ambiguity one has to look back to the history of the language. It is well established that Old Czech did not have any labiodental sounds. In late Middle Ages, /f/ was imported into the language through foreign words, and formed a considerably asymmetrical pair with the already present bilabial approximant /w/. The pull towards greater symmetricity within the system initiated a gradual change of /w/ to /v/. Despite the fact that this change has been going on for over five centuries [1], it has not been completed. Unfortunately, the lack of accurate data prevents us from examining its course. We do not know how fast or steady the process has been and we cannot appreciate the mutual relationship of phonetic and phonological aspects of the change. As indicated above, Czech /v/ keeps some phonological characteristics of a sonorant. As to its phonetic make-up, however, the literature of the past century or so has occasionally questioned its fricativeness on behalf of plosiveness (e.g., [2]), but never put forward the suggestion that the sound may have retained some of its sonorant quality.

Since our informal observations suggested that [v] in contemporary speech is not as unambiguous an obstruent as indicated in literature, we decided to put its phonetic properties to test. The question was whether basic acoustic data fed into a recognition algorithm would place [v] closer to another unequivocal sonorant or obstruent sound. To verify the discriminatory power of our analyses, we also included a pair of related voiceless fricatives, and to address the question of occlusion and release, a plosive sound was added.

Given that we had over 1500 phones uttered by 265 different speakers and the data fed into the recognition algorithms were constrained, it was clear that the recognition rate would be short of 100 %. The resulting errors, then, should provide the answer to our question. If [v] is misclassified as one of the obstruents more often than as the sonorant, then the literature is right in calling it an obstruent without reservations. If, however, it is misclassified as the sonorant more often than as one of the other obstruents, the view of contemporary Czech labiodental continuant should be modified.

2. Method

The material for the present study was extracted from recordings of 265 first-year students at the Faculty of Arts and Philosophy in Prague. The students were asked to read a short story as fluently and as naturally as possible; they were given adequate time for preparation. The recordings were made in a soundproof booth with an electret microphone IMG ECM 2000 and sampled at 22,050 Hz. All further acoustic processing was conducted in the Praat 4.3 software [3].

In choosing the acoustic parameters for our analyses, we were led by universally accepted general differences between the classes of consonants. First of all, it is known that obstruents contain more noise in their spectrum while sonorants are richer in tones. Hence, we took the measure of *harmonicity*. Second, sonorous consonants, especially approximants, are on average shorter than all the others, which prompted us to measure *duration*. Third, due to the nature of different manner of articulation, the *intensity contours* of consonant classes differ in that plosion is manifested as a dynamic bump on the contour, while fricatives have a characteristic dynamic dip before the onset of the following vowel. Approximants have very smooth transitions of their intensity contours into the neighbouring vowels.

Mean harmonicity of the consonants was obtained by means of forward cross-correlation analysis. We measured the mean HNR of the whole segment and, for the voiced consonants, also the mean HNR with the F0 band removed, i.e., filtered from 1.1 times the maximum fundamental frequency in each given segment to Nyquist frequency. Praat’s “undefined” harmonicity values were replaced by -8.4 dB, which corresponds to the mean harmonicity of all voiceless sounds - [s, t, f] - minus three standard deviations.

As to duration, we worked with two variables. The first one was the raw duration of a consonant in milliseconds, the second was the duration normalized to the local articulation rate. This took into account the fact that in Czech, articulation rate drops gradually from the first stress-group of an intonation phrase to the last [4]. Hence, the consonants in the first stress-group have shorter duration than the same consonants in the last stress-group. The nature of our material allowed us to normalize the duration of our consonants by relating it to the duration of the neighbouring vowels:

$$Dur_{norm} = \frac{C_i(dur)}{\frac{1}{2}V_{i-1}(dur) + \frac{1}{2}V_{i+1}(dur)} \qquad (1)$$

On the other hand, the Czech stress pattern does not require normalizing duration with regard to stress since there is no relationship between the two [5].

Intensity measurements were performed on the second half of the consonant and the first half of the following vowel, because that is where the relevant information concerning the manner of articulation was expected (see above). With consonants, we measured the mean intensity of the first third of their second half (1/3 C) and then ten equidistant values in the remaining two thirds of the segment (C1 - C10). With vowels, we measured the mean intensity of the first third (1/3 V) and the mean intensity of the remaining two thirds (2-3/3 V) of their first half. That gave us an intensity contour consisting of 13 values. As we were comparing the intensity of both voiced and voiceless sounds, the frequency range of 0-750 Hz was removed from the sounds for intensity measurements; this frequency range corresponded to fundamental frequency, as well as the first formants in our speech sample. The 13-point intensity contours were compressed by factor analysis with Varimax rotation.

Harmonicity, duration, and intensity-contour data of intervocalic [f, s, t, v, z, l] were subjected to k-means cluster analysis, discriminant analysis and classification-tree analysis [6]. Although we considered using an artificial neural network recognizer, we opted for simpler, but more transparent methods. The same reason led us not to use regular recognition through cepstral or LPC coefficients, because these are sometimes difficult to interpret phonetically.

3. Results

3.1. Recognition of all phones

3.1.1. *Factor analysis of intensity contours*

First of all, factor analysis with Varimax rotation was performed on the intensity data. The originally 13-item dynamic contours were collapsed into three factors. The first three factors explained 93 % of variance in the original data. The first factor (8.10; 62.3 %) relates to the *onset* of the intensity contour, the second factor (2.72; 20.9 %) to its *offset*, and the third factor (1.27; 9.8 %) to the *middle* part. The third factor is the weakest, because there was a lot of overlap in the middle part of the contours. The factor loadings are given in Table 1.

	onset	**offset**	**mid**
1/3 C	**0.927**	0.182	0.133
C1	**0.949**	0.181	0.139
C2	**0.959**	0.165	0.163
C3	**0.944**	0.129	0.228
C4	**0.871**	0.072	0.388
C5	0.687	0.041	0.636
C6	0.420	0.119	**0.862**
C7	0.208	0.294	**0.910**
C8	0.151	0.476	**0.826**
C9	0.182	0.643	0.695
C10	0.216	**0.811**	0.480
1/3 V	0.102	**0.929**	0.287
2-3/3 V	0.189	**0.907**	0.054

Table 1: Factor loadings for intensity contours

The original intensity values were replaced with their respective factor scores, and the new intensity values (onset, offset, mid), together with raw duration in milliseconds and harmonicity of the whole segment, were used in the following analyses.

3.1.2. *K-means cluster analysis*

For the purpose of cluster analysis, the duration and harmonicity data were standardized to z-scores, so as to iron out the differences in their relative magnitudes. The classification matrix in Table 2 shows that cluster analysis was not very successful in discriminating the given phones, with the overall success rate being 77.4 %. The success rate of both [s] and [f] was above 90 %, and neither of them was placed in the /v/-class. Discrimination was considerably worse for the other phones, especially for [z] and [l]. The success rate of [v] was 83.4 %, with most of the incorrectly recognized instances being misplacements in the /l/-class.

	/f/	**/s/**	**/t/**	**/v/**	**/z/**	**/l/**	**success rate [%]**
[f]	247	6	0	0	0	1	97.2
[s]	15	245	0	0	5	0	92.5
[t]	25	26	212	1	0	1	80.0
[v]	1	0	0	221	3	40	83.4
[z]	3	0	0	14	147	84	59.3
[l]	0	0	0	58	58	94	44.8

Table 2: Classification matrix of cluster analysis

3.1.3. *Discriminant analysis*

The classification matrix in Table 3 shows again a high success rate for [s] and [f] (both slightly over 95 %, and neither confused with /v/). The success rate for [t] was 86.4 %, with only one case of confusion with /v/. [v] was successfully discriminated in nearly 90 % of the cases. All but two of the incorrectly discriminated instances of [v] were confusions with /l/. The success rate for [l] and [z] was below

80 %. These two phones were confused with one another, as well as with /v/. The overall success rate of discriminant analysis in recognizing all the analyzed phones is markedly higher than that of cluster analysis - 87.7 %.

	/f/	/s/	/t/	/v/	/z/	/l/	success rate [%]
[f]	243	8	0	0	3	0	95.7
[s]	10	254	0	0	1	0	95.8
[t]	20	14	229	1	1	0	86.4
[v]	1	0	0	238	1	25	89.8
[z]	0	0	0	18	196	34	79.0
[l]	0	0	0	28	21	161	76.7

Table 3: Classification matrix of discriminant analysis

The F-statistic in Table 4 suggests that the two labiodental continuants, [v] and [f], are quite dissimilar. The greatest similarity was revealed between [z] and [l], and also between [v] and [l]. The similarity between [v] and [z] was relatively smaller. All of the other groups differed from each other to a far greater extent.

	[f]	[s]	[t]	[v]	[z]	[l]
[f]		311.0	740.0	948.8	589.7	889.7
[s]	311.0		773.2	1322.3	723.6	1084.4
[t]	740.0	773.2		1245.3	1065.1	1134.7
[v]	948.8	1322.3	1245.3		175.1	86.1
[z]	589.7	723.6	1065.1	175.1		63.0
[l]	889.7	1084.4	1134.7	86.1	63.0	

Table 4: F-statistic of the differences between the classes found in DA

3.1.4. *Classification trees*

The method of discriminant-based linear combination splits turned out to provide the best compromise between simplicity of the tree and the classification success rate.

	/f/	/s/	/t/	/v/	/z/	/l/	success rate [%]
[f]	237	10	3	0	0	4	93.3
[s]	11	254	0	0	0	0	95.8
[t]	11	8	245	0	1	0	92.5
[v]	1	0	0	235	0	29	88.7
[z]	3	3	0	3	201	38	81.0
[l]	0	0	0	14	34	162	77.1

Table 5: Classification matrix of classification trees

The classification matrix in Table 5 shows higher than 90% success rate for [s f t]; moreover, neither of them was confused with /v/. [v] was successfully recognized in nearly 90 % of the cases, and all but one of the incorrectly recognized instances of [v] were confusions with /l/. The success rate for [l] and [z] was again below 80 %. These two phones were confused with one another, as well as with /v/, but - in contrast with discriminant analysis - [z] tended to be confused with /l/ more often, and only exceptionally with /v/. The success rate of classification trees across all the analyzed phones was 88.5 %.

3.2. Recognition of voiced phones only

With nearly 100% certainty, all the previous analyses separate the voiceless phones from the voiced ones. That is why in the second stage we restricted discrimination only to [v l z] with more refined input.

The acoustic data differed from the previous stage in the following manner. The intensity data were compressed again by factor analysis yielding only two factors this time. Instead of using absolute duration of the consonants in milliseconds, the duration of the consonants was normalized against neighbouring vowels (see above). Apart from harmonicity of the phones, we also incorporated harmonicity residue (the amount of harmonicity after fundamental frequency has been filtered out from the segments).

3.2.1. *Factor analysis of intensity data*

Factor analysis with Varimax rotation collapsed the dynamic contour into two factors explaining 94.8 % of the original variance, with the first factor (9.17; 70.5 %) corresponding to the *onset* and the second factor (3.15; 24.3 %) to the *offset*. The factor loadings are shown in Table 6.

	onset	offset
1/3 C	**0.971**	-0.008
C1	**0.982**	0.076
C2	**0.982**	0.136
C3	**0.970**	0.218
C4	**0.939**	0.321
C5	**0.881**	0.445
C6	0.784	0.592
C7	0.646	0.737
C8	0.482	**0.855**
C9	0.320	**0.930**
C10	0.187	**0.962**
1/3 V	-0.006	**0.974**
2-3/3 V	-0.003	**0.847**

Table 6: Factor loadings for intensity contours

3.2.2. *K-means cluster analysis*

If we compare the classification matrix in Table 7 with that of Table 2, we can immediately see a massive improvement in the recognition of [l] and [z] with the new acoustic data. These two segments are confused with one another to a much smaller extent because the harmonic residue for [l] is greater than that for [z] while relative duration of [z] is longer than that of [l].

	/v/	/z/	/l/	success rate [%]
[v]	196	4	65	74.0
[z]	27	212	9	85.5
[l]	29	13	168	80.0

Table 7: Classification matrix of cluster analysis

On the other hand, there is also a considerable deterioration in the recognition of [v] by nearly 10 %. Importantly, most of the incorrectly recognized cases of [v] were again confusions with /l/. Misidentified instances of both [z] and [l] tend to be recognized as /v/ to roughly the same extent. The overall success rate of cluster analysis is again relatively low - 80.0 %.

3.2.3. *Discriminant analysis*

Discriminant analysis provided a result parallel to that of cluster analysis, only with higher success rate. The classification matrix in Table 8 indicates that while the discrimination of [v] remained just below the 90% mark, there was a substantial improvement in the recognition of both [l] and [z] (compare with Tab. 3). The overall success rate of discriminant analysis in classifying all the analyzed phones was 90.6 %.

	/v/	/z/	/l/	success rate [%]
[v]	236	2	27	89.1
[z]	10	227	11	91.5
[l]	15	3	192	91.4

Table 8: Classification matrix of discriminant analysis

The incorrectly discriminated instances suggest that [v] is more similar to [l] than to [z]. This is further confirmed by the F-statistic given in Table 9, which shows that the greatest distance is clearly between [v] and [z] and the smallest distance between [v] and [l], while the distance [l] and [z] is amid the two and relatively larger rather than smaller.

	[v]	[z]	[l]
[v]		430.5	179.5
[z]	430.5		315.7
[l]	179.5	315.7	

Table 9: F-statistic of the classes found in DA

3.2.4. *Classification trees*

The classification matrix in Table 10 shows that the success rate of classification-tree analysis is practically the same as that of discriminant analysis, with all phones lying around the 90% mark. The overall success rate was 89.5 %. Furthermore, comparison with Table 5 shows that the number of instances of correctly discriminated [v] stayed almost the same.

We can also see that while most of the incorrectly recognized instances of [v] were typically identified as /l/, the incorrect placements of [l] and [z] were distributed more evenly among the other two candidates.

	/v/	/z/	/l/	success rate [%]
[v]	234	3	28	88.3
[z]	10	226	12	91.1
[l]	10	13	187	89.0

Table 10: Classification matrix of classification trees

4. Discussion

Classification success rate was the lowest for k-means cluster analysis. The other two types of analyses were about equally successful. The reason for this outcome is that the cluster analysis operated in a five-dimensional space and treated all the variables as having the same importance. Discriminant and classification-tree algorithms were allowed 'to see' the target classes, hence, they could make greater use of the more important features. In analyses with all six phones, both stepwise discriminant analysis and classification trees, for example, revealed that duration and the onset factor were more important than harmonicity and the offset factor, with the mid factor being the least important.

On the other hand, adding the harmonicity residue variable for the discrimination among [v z l] improved the success rate since [l] with removed F0 band retains much more tone than [z]. Similarly, relative rather than absolute durations work better since [l] is clearly shorter than [z] if we neutralize the influence of local articulation rate. However, it is important to notice that even though refining durational and harmonic data leads to greater separation of [l] and [z], it also causes greater confusion of [l] and [v]. This fact helped us to answer the question with which we started our study.

5. Conclusions

All analyses, whether with cruder or more refined representations of harmonicity, duration, and intensity profile, confirmed that modern Czech [v] is in intervocalic position more similar to a sonorant [l] than to any of the other obstruents at hand. Since we used natural connected speech, which inherently entails deviations from canonical articulations (such as frictionless [z]-items, accidental devoicing of segments, or the use of creaky phonation), it would be interesting to see, whether our findings hold for more careful speaking style of contemporary Czech, such as explicit logatom reading, where [v] might be more in line with textbook descriptions.

There are certainly more ways to tackle the problem we have presented. Our aim was to provide a phonetically transparent, yet convincing solution.

6. Acknowledgements

This research was conducted with support of the grant GACR 405/05/0436.

7. References

[1] Lamprecht A., Šlosar D. and Bauer J., *Historický vývoj češtiny*, SPN, Praha, 1977.

[2] Frinta, A., *Novočeská výslovnost*, Praha, 1909.

[3] Boersma, P. and Weenink, D., *Praat, version 4.3.*, www.praat.org, 2005.

[4] Dankovičová, J., *The linguistic basis of articulation rate variation*, Frankurt am Main: Hector, 2001.

[5] Palková, Z., *Fonetika a fonologie češtiny*, Karolinum, Praha, 1997.

[6] StatSoft, Inc., *STATISTICA 7.0*, 2004.