ON THE ROLE OF THE BURST AND TRANSITIONS FOR THE IDENTIFICATION OF PALATALIZED AND NON-PALATALIZED PLOSIVES IN BULGARIAN

Bistra Andreeva , Jacques Koreman & William Barry
University of the Saarland, Institute of Phonetics, Saarbrücken, Germany

ABSTRACT

The effect of palatalization on the articulatory and acoustic realization of Bulgarian plosives is well described. It has been shown in cross-splicing experiments that human listeners are more strongly influenced by the transition than by the burst. In two sets of experiments, palatalization in Bulgarian plosives is investigated in an automatic speech recognition setting. The confusions between phonemes in a plosive identification experiment show that palatalized consonants can be identified well on the basis of closure and burst alone. The addition of vowel transitions leads to a greater improvement in the identification rate for non-palatalized than for palatalized consonants. Follow-up experiments using quasi-cross-splicing indicate that the information carried by the transitions is more important than the information in the closure plus burst for both non-palatalized and palatalized plosives.

1. PHONETIC DESCRIPTION

In Bulgarian, an opposition between non-palatalized and palatalized consonants occurs before back vowels. Tilkov gives a detailed phonetic description of palatalized plosives [1], which are the focus of this paper.

Articulatorily, palatalized plosives are characterized by a raising of the front part of the tongue dorsum to the hard palate in addition to their primary articulation. This results in a vocal tract configuration similar to the vowel /i/. Since the primary articulation for labials (/p, b/) does not involve the tongue, the palatalization is independent of the articulation at the lips. For alveolar plosives (/t, d/) this independence between primary and secondary articulator is reduced, since the tongue is used for both articulations. Palatalization of alveolar plosives leads to a merging of the primary and secondary articulation, affecting the position of the tongue tip and blade, which form a closure low on the alveolar ridge (close to the teeth) for non-palatalized plosives and higher for palatalized ones. Additionally, the blade of the tongue touches the alveolar ridge and the hard palate on both sides of the oral cavity. With velar plosives (/k, g/) the primary and secondary articulations are completely merged, the articulation of palatalized velars being fronted, such that the back of the tongue is raised toward the hard palate, the tongue blade being lowered and the tip touching the lower teeth.

Tilkov writes that for palatalized plosives, the energy at the burst is concentrated in a higher part of the spectrum compared to non-palatalized plosives (see [2], table 1, p. 85, for the shift of the energy concentrations for voiceless plosives). The acoustic result of the secondary raising of the tongue during the palatalized closure phase is that, on release, the oral cavity is divided into two unequal parts, causing lowering of the first and raising of the second formants in the vowel transition (see [2], table 6, p. 96). The table also shows that the vowel onset transitions (F1 and F2 transitions) are longer after palatalized plosives, as is to be expected, with the tongue dorsum changing from a close-front vocoid shape to a back-vowel shape.

In a cross-splicing experiment, Tilkov evaluated the relative information carried by the closure-plus-burst versus the vowel-onset (CV) transitions. He showed that the perception of palatalization mainly depends on the vowel onset transition. By replacing non-palatalized transitions with palatalized transitions, a complete perceptual switch from a non-palatalized to a palatalized plosive could be obtained. An exception to this pattern was the subjects' judgment of cross-spliced velar plosives, for which their decision depended more strongly on the (closure and) burst (see [2], table 8, p. 108).

The work reported here examines whether the relative importance of the vowel transition phase in human perception of palatalized and non-palatalized plosives also applies to machine identification of these sound categories. This can have important implications for ASR systems, because the vowel transitions are not normally exploited for consonant recognition (but see [3])

2. MATERIAL

The stimuli consisted of Bulgarian sentences containing twenty realizations of non-palatalized /p, t, k, b, d, g/ and palatalized /p^j, t^j, k^j, b^j, d^j, g^j/ in different vocalic contexts. It was obviously not possible to fully control these contexts for word stress and vowel quality due to restrictions in the lexicon. The sentences were spoken by 3 male and 3 female speakers and were recorded in a sound-treated room. The speech signal was digitized at 22050 Hz and 12 mel-frequency cepstral coefficients, energy and their corresponding delta parameters were computed from the signal, using HTK [5]. A 15-ms Hamming window was used as a compromise between spectral definition in the analysis and burst duration. The step size was 5 ms and pre-emphasis was 0.97.

All palatalized and non-palatalized plosives and the surrounding vowels were segmented and labelled. The plosives were divided into a closure and a burst (including aspiration).

Voiceless closures, which were labelled with "p0", were segmented on the basis of voicelessness in the signal; for voiced closures, which were called "b0", lack of higher formants was used as the segmentation criterion. Bursts cover the complete duration of the friction. Since the friction may overlap with the beginning of the vowel onset transition, the burst can include voicing. This maximizes the burst duration and minimizes the influence attributable to the vowel onset transition. The burst segments were labelled according to their place of articulation and palatalization (plain "p", "t", "k", "b", "d", "g", vs. palatalized "P", "T", "K", "B", "D", "G"). The vowel offset and onset transitions, with a duration of 35 ms, were automatically derived from the vowel labels. If the vowel was less than 70 ms long, only half the vowel was labelled as a transition. Vowel offset transitions were labelled for example as "I_t" for a transition of "I" to "t", vowel onset transitions as "B_o". The plosives which were investigated were not always preceded by a vowel offset transition.

3. BURST VERSUS TRANSITIONS

In the first set of experiments, the automatic identification of palatalized and non-palatalized plosives was investigated, using hidden Markov modelling. Separate models were trained for the closure phase of the voiced and of the voiceless plosives, for each burst label and for each vowel offset and each vowel onset transition. The hidden Markov models (HMM's) were single mixture 3-state, left-to-right CD-HMM's.

In experiment 1, the plosives were defined in the phoneme dictionary as a sequence of HMM's for the closure and burst, the closure being optional. Thus, the palatalized plosive /B/ was defined as either the HMM for "b0" followed by the HMM for "B", or just the HMM for "B". In the identification test, only the closure and burst of each plosive were cut out from the signal and offered to the system for identification of the consonant.

In experiment 2, the sequence of HMM's defining a plosive was expanded by a vowel offset transition before the closure (e.g. the HMM for "a_B") and a vowel onset transition after the burst (e.g. the HMM for "B_u"). The signal portions offered to the system comprised closure, burst and transitions. The results of the two experiments are shown in tables 1a and b.

	p	t	k	b	d	g	P	T	K	B	D	G	%
p	17	1	5	3	1	2	2	21	17	3	6	24	17
t	3	9	15	1	1	3	2	18	27	4	2	15	9
k	4	3	36	2	1	6	3	13	5	3	3	23	36
b	3	1	10	11	2	2	6	14	22	7	7	16	11
d	6	1	12	4	10	2	4	10	16	7	6	23	10
g	2	2	25	7	3	5	3	14	7	10	7	14	5
P	2	2	17	2	0	2	11	40	8	3	4	8	12
T	1	0	5	0	0	5	1	62	4	3	5	15	62
K	1	3	4	2	0	2	3	32	32	8	4	9	32
B	2	4	9	3	0	3	7	9	7	31	8	17	31
D	2	2	6	3	1	2	5	9	8	13	24	27	24
G	6	7	5	2	1	2	1	4	3	6	8	56	56

Table 1a. Identification rates (rounded percentages) for plosives on the basis of closure and burst only

	p	t	k	b	d	g	P	T	K	B	D	G	%
p	41	5	7	1	8	0	8	13	7	0	2	8	41
t	11	55	6	1	3	1	3	4	10	2	0	3	55
k	3	4	43	3	19	0	4	15	3	1	1	4	43
b	3	6	8	26	11	1	6	13	19	2	3	2	26
d	3	7	7	4	56	1	3	4	9	0	1	5	56
g	3	10	5	2	5	47	4	6	6	4	2	6	47
P	2	11	7	1	7	2	44	17	5	2	1	2	44
T	2	1	3	0	4	3	0	75	3	2	2	5	75
K	2	4	2	1	3	0	6	9	58	3	3	11	58
B	2	5	7	6	10	1	3	10	5	41	3	8	41
D	2	8	8	5	8	6	2	13	6	2	29	11	29
G	1	4	2	3	14	4	1	6	0	3	2	59	59

Table 1b. Identification rates (rounded percentages) for plosives on the basis of closure, burst and vowel transitions

Though well above chance (8.25%), the overall plosive iden-tification rate on the basis of the closure and burst alone is considerably lower (25.69%) than when combined with the vowel transitions (47.91%).

Generally, palatalized plosives can be recognized much more reliably than non-palatalized plosives from the closure plus burst alone. This is particularly true for voiced plosives, which is not very surprising, since the bursts for /b, d, g/ are considerably shorter than for their palatalized counterparts (see also segmentation criteria in section 2). This leads to a poorly defined spectral representation of the burst, making identification intrinsically more problematical. The importance of burst duration for identification is supported by the results for the voiceless non-palatalized velar plosive: the longer average duration of the /k/ burst (see table 2) results in better identification of /k/ than of /p/ and /t/.

phoneme	duration	phoneme	duration
p	16.24	P	27.99
t	19.35	T	42.77
k	36.43	K	54.24
b	8.61	B	16.05
d	11.15	D	28.60
g	19.91	G	27.08

Table 2. Average durations for all plosive bursts (ms)

Comparing identification rates with and without transitions (see table 1b and 1a), we find that the average improvement in identification of non-palatalized plosives is far greater than for palatalized ones. The voiced non-palatalized plosives /b, d, g/, which were identified worst on the basis of closure and burst alone, show the greatest improvement (on average 32.5 percentage points), followed by the voiceless non-palatalized plosives (improvement: 26.6 percentage points).

Without transitions, the percentages correct for palatalized and non-palatalized plosives are 35.9 and 15.2, respectively. If we only consider the correct identification of palatalization, we find 63.5% false alarms for non-palatalized plosives, but only 17.4% for palatalized ones. This indicates the greater relative importance of the burst for the palatalized plosives.

With transitions, plosive identification rises to 51.0% correct for palatalized (an increase of 15.1%) and 44.8% non-palatalized plosives (an increase of 29.6%). This indicates the greater relative importance of the transitions for the non-palatalized plosives. Overall the false alarms also become more evenly balanced (30,5% for non-palatalized vs. 24.9% for palatalized plosives). Whether this greater relative improvement with transitions for non-palatalized plosives is due to the greater importance of the transitions or the lesser importance of the burst information is not clear from these results alone. We shall examine this question experimentally in the following section.

4. QUASI-CROSS-SPLICING

In the first experiment it was shown that it is easier for the system to identify palatalized consonants on the basis of the (closure plus) burst than it was for non-palatalized consonants. The second experiment showed that by adding vowel transitions, the identification of non-palatalized plosives improved more than the palatalized ones. However, we still cannot answer the question whether it is the burst or the transition which is more important for the identification of palatalized and non-palatalized plosives. To answer this question, we carried out another plosive identification experiment, carrying out a simulation of Tilkov's cross-splicing experiment (see section 1).

In that experiment, Tilkov swapped the vowel-onset transition following the palatalized and non-palatalized consonants. However, since the vocalic contexts in our stimulus material are not balanced for palatalized and non-palatalized consonants, it is not possible to replace all the vowel transitions from palatalized plosives by vowel transitions from non-palatalized ones (and vice versa). Therefore, we performed an experiment in which we generalized for place of articulation categories across different vocalic contexts. This was done by mapping the vowel-context-dependent acoustic input onto context-independent place features.

Previous consonant identification experiments [3,4] have shown that consonant identification rates can improve considerably when acoustic parameters are mapped onto phonetic features. In these experiments, phonetic features (like [labial], [alveolar], etc.) were derived from the acoustic parameters by means of a Kohonen network. For vowel transitions, only information relevant to the articulation of the neighbouring consonant was extracted, so that a generalization across vowels is obtained. The phonetic features were subsequently used for hidden Markov modelling.

The same acoustic-phonetic mapping strategy was used in the quasi-cross-splicing experiment presented here. All acoustic parameters (mel-frequency cepstral coefficients, energy and the corresponding delta parameters) were mapped onto the phonetic features [labial], [alveolar], [velar], [palatalized] and [voiced].

These five features were used in a hidden Markov modelling experiment. Models were trained for "p0" and "b0" stop closures, for each of the 12 burst (6 palatalized and 6 non-palatalized bursts, half from voiced and the other half from voiceless plosives), for the 12 (generalized) vowel offset transitions, and for the 12 (generalized) vowel onset transitions. The HMM's modelling the transitions were named V_b, K_V, etc. (where V indicates that the values for the phonetic features were derived from all the different vowels in our material).

To obtain baseline results for the acoustic-phonetic mapping condition, equivalents to experiments 1 and 2 were carried out, but this time using the five phonetic features as input to hidden Markov modelling instead of the acoustic parameters themselves. Since there are fewer HMM's for the vowel transitions, which have now been generalized, the size of the phoneme dictionary (see the beginning of section 3) is much smaller. Overall identification rates are slightly lower than for experiments 1 and 2 (25,08% vs. 25.69% without transitions, and 40.04% vs. 47.91% with transitions).[1] Differentiated according to their non-palatalized vs. palatalized identity, the scores were 17.55% for non-palatalized without and 37.04% with transitions; for palatalized they were 30.57% without and 43.03% with transitions. Since these values merely serve as a baseline for the cross-splicing experiment we shall not go into these results any further.

To imitate Tilkov's cross-splicing experiment with human listeners in an automatic speech recognition (ASR) setting, the phonemes in the phoneme dictionary were redefined: the definition of a non-palatalized plosive was changed so that the non-palatalized vowel transitions were replaced by palatalized. For instance, the phoneme /k/ was defined as a sequence of

$$k = (V\_K)\ (p0)\ k\ K\_V,$$

brackets indicating optionality. The phoneme /K/ is defined, correspondingly as

$$K = (V\_k)\ (p0)\ K\ k\_V.$$

If the burst, which has already been shown to be very important for the identification of the palatalized plosive, is in fact more important than the transitions, a natural realization of /k/ (including natural vowel-/k/ and /k/-vowel transitions) should still be identified as /k/ despite the fact that V_k and k_V belonging to the /K/ entry in the phoneme dictionary. If, however, the transitions are more important than the burst, the presence of a V_k (optional) and k_V transition should lead to identification of /K/.

The results, which are presented in table 3, decisively show that the vowel transitions are more important than the burst, since non-palatalized plosives are mainly identified as their palatalized counterparts, while all palatalized plosives except /K/ are identified more often as their non-palatalized cognates. Correct identification sinks to 13,33% for non-palatalized and to

[1] This can be explained by the loss of redundant information in the acoustic parameters to signal the presence of the five features. The confusions which occur between the phonemes, however, are easier to explain phonetically, the phonetic distance between confused plosive being more frequently the result of a single feature rather than a multiple feature confusion. As a result, the average distance between confused categories is smaller on average (cf. [4] for an evaluation metric of the acoustic distance in the confusion matrices).

19,17% for palatalized plosives, while the misidentification as their cross-palatalized cognates rises to 29.5% for the non-palatalized and to 33.83% for the palatalized plosives. Thus, the combined transitions competing with conflicting bursts achieve higher scores than the bursts alone in the baseline mapping experiment.

	p	t	k	b	d	g	P	T	K	B	D	G
p	17	3	7	8	2	0	34	8	6	6	6	3
t	6	14	3	3	6	1	7	40	10	3	5	2
k	6	7	10	4	3	3	9	9	27	5	3	13
b	6	2	2	12	6	5	10	9	13	26	4	5
d	5	3	6	3	20	2	6	17	4	9	23	2
g	6	3	7	11	3	7	6	8	16	4	3	27
P	39	5	5	5	0	1	16	9	11	7	2	1
T	1	36	4	2	7	3	2	28	8	3	2	4
K	5	3	26	1	2	6	3	14	35	2	3	1
B	16	4	4	35	9	2	4	2	8	9	3	3
D	4	12	6	10	36	3	0	6	2	7	8	6
G	3	3	8	6	7	31	3	5	5	7	4	19

Table 3. Identification rates (rounded percentages) for plosives when both transitions are quasi-cross-spliced

To verify whether it is the vowel onset transition which contributes most to the identification of palatalization (rather than an equal contribution of offset and onset transitions), as one might expect on the basis of the clear formant transitions which are present, a second quasi-cross-splicing experiment was carried out. In this experiment, only the vowel onset (CV) transition was changed from palatalized to non-palatalized and vice versa. In this case, the entry for /k/ in the dictionary is

k = (V_k) (p0) k K_V.

If the information in the vowel onset transition overrules the information in all other parts of the signal, non-palatalized plosives should mainly be identified as palatalized and vice versa, as was the case in the first cross-splice experiment The results show, however, that this is not the case (see table 4).

	p	t	k	b	d	g	P	T	K	B	D	G
p	34	3	4	6	3	3	19	8	8	4	5	3
t	8	25	3	4	6	2	9	27	10	2	5	0
k	8	4	25	9	3	3	7	9	19	3	3	8
b	14	2	6	17	10	6	9	11	9	10	2	2
d	10	8	5	3	27	3	3	13	6	7	11	3
g	10	4	10	11	2	13	2	6	17	6	3	18
P	25	5	5	7	0	0	26	10	11	9	1	1
T	2	25	5	1	8	3	1	39	6	3	2	5
K	6	1	14	4	3	4	3	17	46	0	1	2
B	12	5	2	32	9	2	6	2	10	14	3	4
D	6	6	6	13	28	3	0	11	2	3	16	6
G	3	4	3	9	8	31	2	3	9	6	3	18

Table 4. Identification rates (rounded percentages) for plosives when only the vowel onset transitions is quasi-cross-spliced

The non-palatalized offset-transitions and the non-palatalized burst together achieve a higher recognition rate against a conflicting palatalized onset transition than the burst-alone baseline (23.5% vs. 17.55%.) while the cross-splice cognate rate is comparable to the baseline (17.33%). In contrast, the combined palatalized vowel offset and burst do not achieve the baseline recognition rate in their conflict with non-palatalized onsets (26.5% vs. 30.57%), and the cross-splice cognate rate is also lower than the baseline (25.8%).

5. CONCLUSIONS

In the first experiment, in which plosives were identified on the basis of the closure and the burst, it was shown that the burst carries a lot of information for the identification of palatalized plosives. When vowel offset and vowel onset transitions were added to the signal which was to be identified, the identification rates for non-palatal plosives showed a greater increase than for palatalized ones.

Nevertheless, the combined vowel transitions were clearly decisive for the identification of both non-palatalized and palatalized plosives. This was shown in a quasi-cross-splicing experiment, the simulation within an ASR system of a physical cross-splicing experiment, in which the combined vowel offset and onset transitions, in competition with the burst component, achieved higher recognition scores than the burst alone. A follow-up experiment, in which only the vowel onset transitions in the phoneme dictionary were “cross-spliced” (rather than both transitions), showed that the non-palatalized vowel onset transition is in fact more important for identification than the palatalized onset transition. This indicates that the greater relative improvement to the identification scores for non-palatalized transitions found in experiment 2 (section 3) is not just due to the relatively weak contribution of non-palatalized bursts. The contribution of non-palatalised transitions to identification in competition with palatalization cues confirms that these transitions are more important in absolute terms, despite the articulatorily and acoustically more prominent characteristics of palatalized plosives.

REFERENCES

[1] Tilkov, D. and T. Bojadzhiev. 1977. *Bylgarska Fonetika*. Sofia: Nauka i izkustvo.

[2] Tilkov, D. 1983. Akustichen systav i distribucija na palatalnite syglasni v knizhovnija bylgarski ezik. In Bojadzhiev , T. and M. Mladenov (eds.), *Izsledvanija vyrhu bylgarskija ezik*, pp. 79-139. Sofia: Nauka i izkustvo.

[3] Koreman, J., W.J. Barry and B. Andreeva. 1998. Exploiting transitions and focussing on linguistic properties for ASR. *Proceedings ICSLP’98*.

[4] Koreman, J., B. Andreeva and W.J. Barry. 1998. Do phonetic features help to improve consonant identification in ASR? *Proceedings ICSLP’98*.

[5] Young, S., J. Jansen, J. Odell, D. Ollason and P. Woodland. 1995. *The HTK Book*. Cambridge: Cambridge University.