On the Impact of Labialization Contexts on Unit Selection Speech Synthesis

Daniel Tihelka[1], Zdeněk Hanzlíček[1], Pavel Machač[2], Radek Skarnitzl[2] and Jindřich Matoušek[1]

[1]Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic
[2]Institute of Phonetics, Charles University in Prague, Czech Republic

{dtihelka,zhanzlic,jmatouse}@kky.zcu.cz
{pavel.machac,radek.skarnitzl}@ff.cuni.cz

***Abstract*— *This paper presents a study on coarticulatory labialization and the significance of its respecting/violation during selection and concatenation of speech units in the unit selection speech synthesis. The aim of this study is to improve the overall speech quality, especially to increase the perceptual inconspicuousness between concatenated units. The labialization importance was verified by two listening tests—for phonetic laymen and specialists. To suppress the influence of other factors, both tests contained utterances with specially selected phones in specific contexts with respected and violated labialization. The preference for items with correct labialization was evident, which confirms the benefit of considering coarticulatory labialization in a unit selection speech synthesis.*

Keywords*— *coarticulatory labialization, speech synthesis, unit selection.

I. Introduction

Phonetic context plays an important role in a concatenative speech synthesis since it affects correct transitions of spectral features of concatenated units. Despite this fact, its handling is usually somewhat limited to direct matching of a particular context phone [1] or is indirectly modelled by a data-driven technique, e.g. clustering [2], driven by various acoustic features.

After purely phonetically motivated approaches (like formant synthesis, see e.g. [3]) and purely data-driven approaches (like unit selection [4], [5] or statistical parametric speech synthesis [6]) have almost exhausted their options, combination of both phonetic knowledge and data-driven, or statistical processing starts to be researched (for instance as in [7], [8], [9]), suggesting that deeper incorporation of phonetic knowledge into speech technology, in the appropriate form, is necessary (a successful integration of phonetic features in concatenative speech synthesis was also reported in several studies, e.g. [10], [11]). This paper follows these tendencies. More specifically, it deals with the unit-selection approach and, exploiting explicit phonetic knowledge, it aims mainly at removing the disruptive effects caused by mismatches in labialization contexts from synthetic speech.

This paper is organized as follows. Section II introduces the theoretical phonetic background of the researched phenomenon. Section III deals with the text-to-speech system employed in our research. The main attention is focused on the utilization of the phonetics features within the unit selection process. Experiments evaluating the importance of respecting the labialization are described in Section IV and their results are presented in Section V. Finally, Section VI summarizes the paper and outlines our future work.

This research was supported by the Technology Agency of the Czech Republic, project No. TA01030476. The access to the METACentrum clusters provided under the research intent MSM6383917201 is highly appreciated.

978-1-4673-5604-6/12/$31.00 ©2012 IEEE

II. The role of phonetic contexts during concatenation

In natural speech, the physiologically conditioned incomplete synchronization of articulatory and phonatory gestures leads to coarticulation, the mutual influencing of neighbouring speechsounds. An articulating organ may anticipate the position of the following speechsound (regressive coarticulation), remain in the position of the preceding speechsound (progressive coarticulation), or anticipate and remain at the same time (fusional coarticulation). The articulatory overlaps may then affect the phonetic neighbourhood (i.e., a part of the neighbouring speechsound, the entire speechsound or even more of them). Apart from coarticulatory labialization, on which this research is focused, we may mention nasalization, which results from lowering the velum already during the vowel preceding the nasal consonant.

Based on the concept of phonetic features [12], [13], we talk about the extension of some of the *inherent phonetic features* of one speechsound (e.g., the nasality of a nasal consonant) on another speechsound, in which it operates as an *extrinsic phonetic feature*. This extrinsic feature will result in a difference from the canonical makeup of the speechsound (nasalization of an oral speechsound). Such "colouring" of the sound tends not to be phonologically distinctive, and an ordinary language user cannot perceive the presence of extrinsic phonetic features.

However, the presence of coarticulation may lead to problems in concatenative speech synthesis: when one part of a speechsound has no coarticulatory features and the second part is "enriched" with a coarticulatory conditioned extrinsic feature, we may predict a higher probability of such concatenation to have an intrusive effect.

Our present research focuses on coarticulatory labialization and its impact in unit concatenation. One of the inherent phonetic features of back vowels in Czech is labialization

which may, as an extrinsic feature, spread to a neighbouring consonant and thus affect their phonetic makeup. From the acoustic perspective, the main consequence of labialization is the lowering of F_2 and partly also F_3. Other Czech vowels are not labialized and may be classified as belonging to one group.

Let us take the example of the laryngeal fricative /h/ (which is, in Czech, voiced: [ɦ]). We expect a high probability of coarticulatory effects in this speechsound, since the supraglottal articulating organs do not actively contribute to its sound shape, and these organs can therefore assume the positions pertaining to the neighbouring speechsounds. Assuming the existence of mainly regressive coarticulation in Czech, if /h/ is followed by a back vowel, the /h/ itself will be partly or entirely realized with rounded lips. If a word is chosen from the database in which the presence or absence of labialization in the phonetic context of /h/ will not be in agreement with natural coarticulation, the likelihood of an intrusive effect of the concatenated /h/ is increased. Four possible labialization contexts of an intervocalic consonant are listed in Table 1 where $\mathrm{V}^{+}_{\mathrm{lab}}$ denotes a labialized vowel, $\mathrm{V}^{0}_{\mathrm{lab}}$ an unlabialized vowel and C a consonant.

Table 1. *Labialization contexts of intervocalic consonants.*

	Context	Example	
(1)	$\mathrm{V}^{0}_{\mathrm{lab}} - \mathrm{C} - \mathrm{V}^{0}_{\mathrm{lab}}$	*vyhynul*	[v**ɪ**ɦ**ɪ**nul]
(2)	$\mathrm{V}^{+}_{\mathrm{lab}} - \mathrm{C} - \mathrm{V}^{+}_{\mathrm{lab}}$	*v kruhu*	[fkr**uɦu**]
(3)	$\mathrm{V}^{0}_{\mathrm{lab}} - \mathrm{C} - \mathrm{V}^{+}_{\mathrm{lab}}$	*v lihu*	[vl**ɪɦu**]
(4)	$\mathrm{V}^{+}_{\mathrm{lab}} - \mathrm{C} - \mathrm{V}^{0}_{\mathrm{lab}}$	*v kruhy*	[fkr**uɦɪ**]

The natural "colouring" of the consonant due to coarticulatory labialization can differ greatly in similar contexts, especially when comparing conditions (1) and (4) from Table 1. For example, it would be best to synthesize the /h/ in the word *v**yhy**nul* from words in which the labialization context of /h/ is in agreement, i.e. $\mathrm{V}^{0}_{\mathrm{lab}} - \mathrm{C} - \mathrm{V}^{0}_{\mathrm{lab}}$. If, on the other hand, the labialization context is violated, and the word is synthesized from the words *l**ih**em* (correct) and *u**hy**bat*, we predict a higher probability of discontinuity in the concatenation point, caused by incongruent labialization.

The aim of this experiment is to verify perceptually the hypothesis of the intrusive effect of coarticulatory labialization, and subsequently to formulate penalization rules for the automatic selection of words from the database.

III. Concatenation in the ARTIC TTS system

Regarding to features used during the selection process, the current setting of our TTS system ARTIC [14] is somewhat of a hybrid between the "classic" concept usually used in today's unit selection frameworks, where feature values are compared directly with *equal/different* result, and (non-discrete) "suitability" [15] which intends to move from simple match/mismatch features comparison forward to what we call *prosodic synonymy/homonymy* of units [16].

It is also the *phonetic context*, as the sub-feature of the whole target feature vector, still following the classic concept. To be more precise, let *vyhynul* [vɪɦɪnul] be the text to be synthesized. It consists of the following sequence of diphones, with left and right context shown as lower index left to or right to the diphone ($*$ stands for *not important for the example*):

$$ {}_{*}\mathrm{v}\mathrm{ɪ}_{\mathrm{ɦ}} \quad {}_{\mathrm{v}}\mathrm{ɪ}\mathrm{ɦ}_{\mathrm{ɪ}} \quad {}_{\mathrm{ɪ}}\mathrm{ɦ}\mathrm{ɪ}_{\mathrm{n}} \quad {}_{\mathrm{ɦ}}\mathrm{ɪ}\mathrm{n}_{\mathrm{u}} \quad {}_{\mathrm{ɪ}}\mathrm{nu}_{\mathrm{l}} \quad {}_{\mathrm{n}}\mathrm{ul}_{*} $$

When a diphone [ɪɦ] is to be synthesized, its required left and right contexts (among other features) are matched against the *real* contexts of all the candidates of the diphone (i.e. the context they were recorded in). The context-related target subcost is equal to 0 only if both contexts match and greater than 0 otherwise, no matter the actual difference of phones in the context. In this way, the context mismatch penalty is the same for [n/m] and [n/t] mismatches (with the [*required*/*have*] ordering). It means that if the unit selection process is unable to ensure the right match of both contexts, e.g. there is no exact context available or the context sub-feature match is "sacrificed" in favour of matching more important features, the sequence of the diphones $[{}_{*}\mathrm{v}\mathrm{ɪ}_{\mathrm{o}}\ {}_{\mathrm{u}}\mathrm{ɦ}\mathrm{ɪ}_{*}]$, with the real context as shown, may be chosen to build the synthetic speech.

Although the most prominent coarticulation effect is embedded into each diphone, the deflection of what should be naturally pronounced may also influence (progressively in case of left, or regressively in case of right context) the phone in which the concatenation is occurring. As a result, phone halves with different "colouring" can be joined together, which may result in audible unnatural artifact.

To support the claim, we have looked at the context feature mismatch in the synthesis of 5,000 randomly selected phrases, consisting of 195,964 diphones being concatenated. The result is in Table 2, where *left* and *right* rows display mismatch of the particular context not taking into account the other, while *both* row displays mismatch of both contexts at the same time.

Table 2. *The mismatch of context sub-feature in 195,964 diphones concatenated. The column* Labialization *shows the number of contexts with incongruent labialization, see Section IV.*

Context	Mismatched	%	Labialization	%
left	46,915	23.9	8490	4.3
right	46,683	23.8	7975	4.1
both	11,014	5.6	×	×

It can be seen that there is the insignificant number of units is used with mismatched context. Considering that there is no way of distinguishing mismatch significance, the current handling of context feature only, in fact, increases the target cost value, without actually improving the synthetic speech[1]. Therefore, we define the following requirements the correct behaviour should follow:

[1]Let us emphasize that this is general problem not only of the context feature, but of every particular feature measuring a mismatch on the same principle!

- Mismatch should be penalized only if it may, in theory or according to a trained setting, lead to audible effects. For the context it means to avoid such interchange which is expected, according to the phonetic theory, to affect the "colouring" of the speechsound.
- There must be several levels of mismatch impact (not necessarily discrete), ranging from acceptable to forbidden, which may also be cross-feature dependent (e.g. mismatch of prosodeme *P0* and *Px* [17] may be acceptable for *Px* beginning position, but forbidden on *Px* end). The leveling acts as counterbalance of feature importance weighting, allowing to use acceptable mismatch even for a feature with higher importance in favor of features with not so acceptable mismatch.

Our (long term) aim is to shift the role of features (not only the context-related one, but all in general) from prescribing what is required to avoiding what is not desired (possibly causing problems for the given positioning in the synthesized text). Many our observations indicate that it has large potential to increase the quality of unit selection generated speech—especially in reducing audible artifacts and lowering the dependency on the size of corpus.

IV. Speech synthesis experiments with labialization contexts

Although there are only 4% of context mismatches affecting the labialization, it does not mean that the remaining 20% of context mismatch can be ignored. There may be other problems, e.g. nasalization, as well as there may be cases where the interchange is not expected to cause any problems and it only unnecessarily increases the target cost (forbidding closer match of other features). However, it is left to be addressed in our future work.

In the present paper we limit ourself to the examination of labialization effect, since according to the phonetic theory the labialization congruence is supposed to be clearly audible. Also, we use specially designed test stimuli, since:

1) we need to prove if the phonetic theory we base our assumptions on is valid;
2) we need to prove the importance of labialization congruence in synthetic speech, especially if its violation is perceived by ordinary TTS users (non phonetic experts) as unnatural,
3) but at the same time we want to avoid secondary effects, not related to the examined phenomenon, causing unnatural artifacts (glitches, etc.), since they would distort the results of listening tests.

The complex labialization situation for particular phones is denoted by $\mathcal{L}(L, R)$ where L and R describe the labialization congruence of the left and right phones, respectively. L and R are composed of the combination of 2 symbols $+$ (labialized, see Table 1) and 0 (not labialized); the first symbol represents the context type required to be synthesized, the second symbol is the *real* context type of the examined consonant phone (i.e. *required*/*have* ordering). Thus, the congruent labialization is denoted "++" or "00" and incongruent "0+" or "+0". For congruent labialization of both contexts, i.e. $\mathcal{L}(00, 00)$, $\mathcal{L}(00, ++)$, $\mathcal{L}(++, 00)$ and $\mathcal{L}(++, ++)$, a simplified general notation was introduced $\mathcal{L}(=, =)$. Similarly, notation for partly or fully incongruent labialization was defined as $\mathcal{L}(=, \neq)$, $\mathcal{L}(\neq, =)$, or $\mathcal{L}(\neq, \neq)$.

The labialization, as described in Section II affects *vocal-consonant-vocal* (VCV) phones sequence. However, the ARTIC works internally with diphones, so the VCV sequence is synthesized by two diphones VC and CV. For the diphones, we are interested in labialization congruence of right context of ${}_{*}\mathrm{VC}_{\mathrm{V}} = L$ and the left context of ${}_{\mathrm{V}}\mathrm{CV}_{*} = R$. To illustrate the notation, let us require a diphone sequence [$_{*}$ɪfiu $_{ɪ}$fiu$_{*}$] for synthesis, having candidates [$_{*}$ɪfi$_{\mathrm{o}}$], [$_{*}$ɪfi$_{\mathrm{r}}$], [$_{*}$ɪfi$_{\mathrm{a}}$], and [$_{\mathrm{a}}$fiu$_{*}$], [$_{\mathrm{o}}$fiu$_{*}$], [$_{\mathrm{r}}$fiu$_{*}$] to be selected from. Obeying the labialization principle, the use of individual candidates will lead to the following congruency marking:

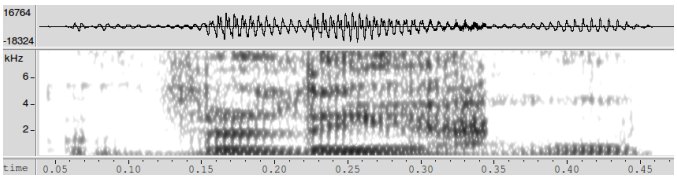
$$[{}_{*}\text{ɪfi}_{\text{u}}\ {}_{\text{ɪ}}\text{fiu}_{*}] \mapsto [{}_{*}\text{ɪfi}_{\text{o}}\ {}_{\text{a}}\text{fiu}_{*}] \ldots \mathcal{L}(++, 00) \ldots \mathcal{L}(=, =)$$
$$[{}_{*}\text{ɪfi}_{\text{u}}\ {}_{\text{ɪ}}\text{fiu}_{*}] \mapsto [{}_{*}\text{ɪfi}_{\text{a}}\ {}_{\text{r}}\text{fiu}_{*}] \ldots \mathcal{L}(+0, 00) \ldots \mathcal{L}(\neq, =)$$
$$[{}_{*}\text{ɪfi}_{\text{u}}\ {}_{\text{ɪ}}\text{fiu}_{*}] \mapsto [{}_{*}\text{ɪfi}_{\text{r}}\ {}_{\text{o}}\text{fiu}_{*}] \ldots \mathcal{L}(+0, 0+) \ldots \mathcal{L}(\neq, \neq)$$

Let us emphasize that $++$ and 00 cases *do not* mean the exact match of context diphones! It only means that the context phones used do not violate the labialization assumption, although they differ (as shown in the example above).

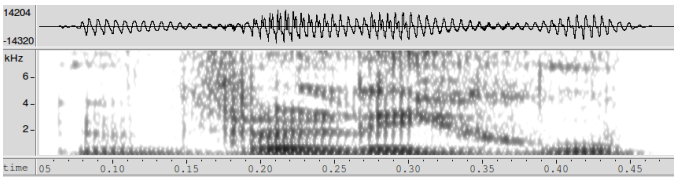
In the experiment, we have selected several words with phones [l], [j], [fi], [x], [s], [z], [r], and [v] in the VCV contexts; during preliminary listening tests, these consonants manifested higher susceptibility to coarticulatory labialization. We let the real TTS system ARTIC [14] to synthesize the words, using male voice [18], and the variants without unnatural artifacts not related to the examined diphones—those on which the labialization effect is expected, were chosen. For those variants, we have limited the set of candidates of each examined diphone independently to those which have the appropriate left and right context phones to either follow (00, $++$) or violate ($+0$, $0+$) labialization principle (the other context, marked by $*$ in the example above, represented an arbitrary phone except for a nasal). For each particular setting, the words were synthesized again, using the limited set of candidates to be chosen from. Note that except the candidates set limitation, the selection algorithm itself has not been modified in any way. The result of synthesis for $\mathcal{L}(\neq, \neq)$ (both left and right contexts violate labialization) and $\mathcal{L}(=, =)$ (both contexts obey labialization) settings is illustrated by Figure 1.

V. Results and discussion

The effect of respecting or violating the labialization context in the unit selection speech synthesis was assessed by performing two preference listening tests—for listeners with and without a phonetic education. As a side effect, it should illustrate the influence of knowledge of the phonetic background on the perception and rating of the researched phenomenon. Both



(a) violated labialization context $\mathcal{L}(\neq,\neq)$ of consonant [fi] when synthesized [ɪfiu ɪfiup] by diphones [ʒɪfia: ofiup]



(b) respected labialization context $\mathcal{L}(=,=)$ of the same consonant [fi] as above by diphones [jɪfiu ɪfiuf]

Fig. 1. Waveforms and spectrograms of the synthetized phrase *"v lihu"* [vlɪfiu].

tests contained a set queries composed of pairs of utterances. The following simple scale was used for evaluation:

1) Utterance A sounds better than utterance B.
2) Both utterances are perceptually similar.
3) Utterance A sounds worse than utterance B.

A. Listening test for non-phoneticians

The first listening test was intended for phonetic laymen. 19 listeners without phonetic education took part in this test, most of them had some former experiences with synthetic speech and listening tests. Test contained 40 queries. Only combinations of fully respected $\mathcal{L}(=,=)$ and fully violated $\mathcal{L}(\neq,\neq)$ labialization contexts were compared.

Table 3. *Preference test for non-phoneticians [%].*

Preferred $\mathcal{L}(\neq,\neq)$	No preference	Preferred $\mathcal{L}(=,=)$
19.5	25.1	55.4

In accordance with our predictions, lay listeners preferred items with the respected labialization context over items with the violated labialization context. However, in several queries, a significant preference for items with violated labialization context $\mathcal{L}(\neq,\neq)$ was noticed. Despite the strictly controlled experiment, an additional inspection of those items mostly revealed some minor artifacts in utterances with respected labialization context $\mathcal{L}(=,=)$. The influence of these artifacts apparently exceeded the significance of labialization congruence.

B. Listening test for phoneticians

Participants of the second listening test were 8 students of phonetics. They were informed about the phenomenon and instructed to try to disregard other potential intrusive features. Test contained 112 queries; all combinations of respecting and violating the labialization context were compared with the exception of phones [s], [z], [r] and [v] which were compared only with ++ or 00 labialization context.

Table 4. *Results of preference test for phoneticians. Each cell contains 3 values [%]: preference of labialization corresponding to the row, no preference, preference of labialization corresponding to the column.*

	$\mathcal{L}(\neq,\neq)$	$\mathcal{L}(\neq,=)$	$\mathcal{L}(=,\neq)$	$\mathcal{L}(=,=)$
$\mathcal{L}(\neq,\neq)$	×	10.9 25.0 64.1	15.6 28.1 56.3	10.6 15.2 74.2
$\mathcal{L}(\neq,=)$	64.1 25.0 10.9	×	44.5 21.1 34.4	26.5 26.6 46.9
$\mathcal{L}(=,\neq)$	56.3 28.1 15.6	34.4 21.1 44.5	×	18.0 25.8 56.2
$\mathcal{L}(=,=)$	74.2 15.2 10.6	46.9 26.6 26.5	56.2 25.8 18.0	×

In agreement with the test for lay listeners, items with respected labialization context $\mathcal{L}(=,=)$ were preferred over all types of partly or fully violated labialization context, i.e. $\mathcal{L}(\neq,=)$, $\mathcal{L}(=,\neq)$ and $\mathcal{L}(\neq,\neq)$. Moreover, the preference for items with correct labialization was even more decided than in the first test.

By analyzing the comparisons involving the combination $\mathcal{L}(=,\neq)$ and $\mathcal{L}(\neq,=)$ we can conclude that the respected labialization context of the right diphone (i.e. regressive labialization) seems to be slightly more perceptually important than the respected labialization context of the left diphone (i.e. progressive labialization). This inference is supported by the direct comparison of $\mathcal{L}(=,\neq)$ and $\mathcal{L}(\neq,=)$ and also by their indirect comparison by using $\mathcal{L}(=,=)$ and $\mathcal{L}(\neq,\neq)$.

Table 5 presents the comparison of respected and violated labialization for individual consonants. Our assumption on labialization significance is confirmed here as well, because respected labialization contexts were preferred in most cases for all the compared consonants, while violated labialization contexts are preferred significantly less. Among phones [x], [fi], [j] and [l], which were compared in more phonetic contexts, the velar fricative [x] seems to be affected the most by coarticulatory labialization, and the lateral approximant [l] the least. However, more experiments would be necessary to support such detailed results.

Table 5. *Comparison of fully respected and fully violated labialization for the selected consonants.*

Prefer.	[x]	[fi]	[j]	[l]	[s]	[z]	[r]	[v]
$\mathcal{L}(\neq,\neq)$	5.6	24.4	16.9	18.8	25.0	0.0	6.3	31.2
None	25.0	16.9	26.9	28.4	25.0	6.3	18.8	25.0
$\mathcal{L}(=,=)$	69.4	58.8	56.3	52.8	50.0	93.8	75.0	43.8

The comparison of partly violated labialization contexts $\mathcal{L}(\neq,=)$ and $\mathcal{L}(=,\neq)$ for individual phones is presented in

Table 6. Here, our prediction that respecting regressive labialization will be preferred to respecting progressive labialization has been confirmed for three out of the four consonants: [j] is showing the opposite tendency. Again, more experiments and a thorough analysis is necessary to verify this conclusion.

Table 6. *Comparison of partly violated labialization contexts for particular consonants.*

Preference	[x]	[fi]	[j]	[l]
$\mathcal{L}(\neq,=)$	40.6	53.1	34.4	50.0
None	34.4	12.5	18.8	18.8
$\mathcal{L}(=,\neq)$	25.0	34.4	46.9	31.3

C. Consistency in rating

To test the consistency in rating of particular listeners during the listening test, several pairs of utterances were randomly repeated. Following types of consistency can be distinguished:

- *Full consistency* – the rating was equal in both cases, i.e. the same item (or none) was preferred.
- *Semi consistency* – in one case, one of items was preferred, in the other case, both were evaluated as similar.
- *Zero consistency* – the rating was contradictory.

For a better comparison of the consistency of both listening tests, only pairs composed of types $\mathcal{L}(\neq,\neq)$ and $\mathcal{L}(=,=)$ were repeated (the first test contained only these combinations). In all test the same number of queries (7) was repeated. The results are presented in Table 7.

Table 7. *Consistency in rating [%].*

Type of consistency	Full	Semi	Zero
Phoneticians	83.9	12.5	3.6
Non-phoneticians	75.2	17.3	7.5

Although the test for phoneticians was longer and more difficult, their rating was more consistent. Moreover, the 100%-consistency was reached by 4 of 8 phoneticians but only by 2 of 19 lay listeners.

VI. Conclusion and future work

In this paper, a study on the importance of considering coarticulatory labialization within the unit selection speech synthesis was presented. Although the experiment was carried out under rather controlled conditions, the results clearly confirm the initial hypothesis that the violation of coarticulatory labialization during unit concatenation may lead to intrusive effects on listeners, even those without phonetic background and trained hearing. The statistical significance of that conclusion was confirmed by using the sign test; the p-value was lower than 0.01 for both listening tests.

In our future work, we plan to incorporate findings gained from the described experiments in our TTS system. Probably some additional experiments will have to be performed to specify the importance of labialization for particular phones in specific contexts and to set the proper penalty weighs. The performance of the default and modified TTS systems will be compared by using ordinary sentences where more various factors can affect the overall quality of resulting speech. The influence of other phonetic features (e.g. nasality) will also be targeted.

In any case, our findings assure us that it is beneficial to formulate (and/or train, based on real data of a particular speaker) *substitution penalty matrix* for unit selection which will account for the natural coarticulatory phenomena. But even more importantly, they assure us that unit selection is supposed to perform better when the features are set to avoid what should not appear in the synthetic speech rather than trying to follow a (possibly inaccurate) target specification closely.

References

[1] Clark, R., Richmond, K., King, S., "Festival 2 – Build Your Own General Purpose Unit Selection Speech Synthesizer", in: Proceedings of SSW 2004, 173–178, Pittsburgh, USA, 2004.

[2] Black, A., Taylor, P., "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", in: Proceedings of Eurospeech 1997, 601–604, Rhodes, Greece, 1997.

[3] Allen, J., Hunnicutt, M. S., Klatt, D., "From Text to Speech: The MITalk System", Cambridge University Press, 1987.

[4] Hunt, A., Black, A. W., "Unit Selection in Concatenative Speech Synthesis System Using a Large Speech Database", in: Proceedings of ICASSP 1996, 373–376, Atlanta, USA, 1996.

[5] Clark, R., Richmond, K., King, S., "Multisyn: Open-Domain Unit Selection for the Festival Speech Synthesis System", in: Speech Communication, vol. 49, 317–330, 2007.

[6] Zen, H., Tokuda, K., Black, A. W., "Statistical Parametric Speech Synthesis", in: Speech Communication, vol. 51, 1039–1064, 2009.

[7] Barry, W. J., van Dommelen, W. A. (Eds.), "The Integration of Phonetic Knowledge in Speech Technology", Heidelberg: Springer, 2005.

[8] van Santen, J., "Phonetic knowledge in text-to-speech synthesis", in: Barry, W. J., van Dommelen, W. A. (Eds.), The Integration of Phonetic Knowledge in Speech Technology, 149–166, Heidelberg: Springer, 2005.

[9] Black, A. W., Bunnel, T., Dou, Y., "Articulatory Features for Expressive Speech Synthesis", in: Proceedings of ICASSP 2012, 4005–4008 Kyoto, Japan, 2012.

[10] Kawai, H., Tsuzaki, M., "Acoustic Measures vs. Phonetic Features as Predictors of Audible Discontinuity in Concatenative Speech Synthesis", in: Proceedings of Interspeech 2002, 2621–2624 Denver, USA, 2009.

[11] Syrdal, A., Conkie, A., "Perceptually based Data-driven Join Costs: Comparing Join Types", in: Proceedings of Interspeech 2005, 2813–2816, Lisbon, Portugal, 2005.

[12] Machač, P., "Stabilita zvukových charakteristik fonémů ve spontánních mluvených projevech", in: Z. Hladká, P. Karlík [Eds], Čeština univerzália a specifika, 5:427–435, Nakladatelství Lidové noviny, Praha, 2004. (*in Czech*)

[13] Machač P. and Skarnitzl R., "Principles of Phonetic Segmentation", Epocha, Prague, 2009.

[14] Matoušek, J., Tihelka, D. and Romportl, J., "Current state of Czech text-to-speech system ARTIC", in: Text, Speech and Dialogue, Lecture Notes in Computer Science, 4188:439–446, Springer, Berlin, Heidelberg, 2006.

[15] Tihelka, D. and Matoušek, J., "Unit selection and its relation to symbolic prosody: a new approach", in: Proceedings of Interspeech 2006, 2042–2045, Pittsburgh, USA, 2006.

[16] Tihelka, D. and Romportl, J., "Exploring Automatic Similarity Measures for Unit Selection Tuning", in: Proceedings of Interspeech 2009, 736–739, Brighton, England, 2009.

[17] Romportl, J:, Matoušek, J., "Formal Prosodic Structures and their Application in NLP", in: Text, Speech and Dialogue, LNCS, 3658:371–378, Springer, Berlin, Heidelberg, 2005.

[18] Matoušek, J, Romportl, J:, “Recording and annotation of speech corpus for Czech unit selection speech synthesis”, in: Text, Speech and Dialogue, LNCS, 4629:326–333, Springer, Berlin, Heidelberg, 2007.