The Role of Nasal Contexts on Quality of Vowel Concatenations

Milan Legát[1] and Radek Skarnitzl[2,*]

[1] University of West Bohemia in Pilsen, Faculty of Applied Sciences
Department of Cybernetics, Czech Republic
legatm@kky.zcu.cz
[2] Charles University in Prague, Faculty of Arts
Institute of Phonetics, Czech Republic
radek.skarnitzl@ff.cuni.cz

Abstract. This paper deals with the traditional problem of occurrence of audible discontinuities at concatenation points at diphone boundaries in the concatenative speech synthesis. We present results of an analysis of effects of nasal context mismatches on the quality of concatenations in five short Czech vowels. The study was conducted with two voices (one male and one female), and the results suggest that the female voice vowels /a/, /e/ and /o/ are inclined to concatenation discontinuities due to nasalized contexts.

Keywords: speech synthesis, unit selection, concatenation cost, nasality, phase mismatch, pitch marks.

1 Introduction

Despite the increasing popularity of HMM based and hybrid speech synthesis methods, the unit selection concatenative systems still represent the mainstream in many real life applications, especially in limited domains where synthesized chunks are combined with pre-recorded prompts. In such applications, the ability of the unit selection to deliver highly natural and to the recordings well fitting output are the key factors. Not surprisingly, the unit selection also remains the first choice for eBook reading applications, which have been acquiring a lot of interest over the recent years.

Among the unit selection related issues that continue to be non-resolved, the audible discontinuities appearing at concatenation points play an important role. Many studies have been published over the last one and a half decades, dealing with the design of concatenation cost functions [1,2,3], to name but a few. Despite the considerable amount of efforts, none of them unfortunately succeeded in providing a clear answer on how to measure the concatenation discontinuities. The presented results have even sometimes been in contradiction.

A few papers, addressing the issue of phonetic effects on vowel concatenations [4], phonetic effects on discontinuity detection in general [5], or using phonetic features as a support for acoustic measures [6], have also been published.

* This research was supported by the Technology Agency of the Czech Republic, project No. TA01030476, and by the grant of the University of West Bohemia, project No. SGS-2010-054.

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 551–558, 2012.

We believe that the main contribution of this paper consists in a detailed analysis of one particular example of concatenation artifacts, which shows how complex the problem is, and that generalizations are difficult to make in this area. This could be one of the explanations for the failing of most of the traditional concatenation cost functions, whose common feature is speaker and phoneme generalization, to reliably reflect the human perception of the concatenation artifacts.

The rest of this paper is organized as follows. The next section gives a motivation for choosing nasals as the context of interest of this study. Section 3 describes our experimental set up, including a data collection and a listening test design. In Section 4, we present the results of our experiment and form a hypothesis to explain the obtained results. Section 5 describes more detailed analysis of the concatenation points with respect to the hypothesis, and finally, in Section 6, we draw conclusions and outline some future work intentions.

2 Motivation

From the articulatory perspective, nasalized vowels are quite simple—they differ from their oral counterparts only in lowering the velum. However, a large and complex resonance space emerges as a result of opening the nasal cavity, which is why nasalized vowels and vowels in the context of nasals in general represent, from the acoustical perspective, probably the most complicated sounds of human speech. This simple articulatory gesture is acoustically manifested in various ways, depending especially on the quality of the vowel and also on the degree of acoustic coupling between the two resonance chambers. For these reasons—and also because the nasal cavity and the paranasal cavities of every speaker are different—there are only few universal acoustic correlates of nasality.

The acoustic complex of nasalized vowels consists of nasal formants (formants of the pharyngonasal tract), oral or vocalic formants (whose frequency may, however, be shifted compared to non-nasal vowels), and antiformants which frequently appear in pairs with nasal formants [7].

The most important acoustic features responsible for the perceptual impression of nasality appear in low frequencies. One of the main correlates of nasality, regardless of vowel quality, is the relative lowering of the intensity of F1, specifically by 6–8 dB according to [9]. The second "universal" feature related to the nasality is the presence of a spectral peak around 250 Hz, which corresponds to the first formant of the pharyngonasal tract, typically marked as N1. The presence of antiformants due to coupling of the nasal cavity is also universal, but their specific frequencies differ in various studies (see [10] for a review).

For the purposes of concatenating units in a speech synthesis system, the presence of nasality in only one of two concatenated diphones may lead to perceived discontinuities, which was indeed observed in our informal experiments. A nasalized vowel may, on the one hand, manifest higher intensity in low frequencies (around 250 Hz, the N1) and, on the other hand, the energy roll-off above this peak is likely to be considerably greater due to the weaker F1 and generally stronger spectral slope. Our hypothesis is that it may be undesirable to concatenate a nasalized vowel with a nonnasal vowel or vice

versa, since the energy difference in specific frequency bands may cause the impression of discontinuity. The aim of this paper is thus to examine whether controlling for the contextual nasality conflicts will lead to better continuity of concatenations in vowels.

3 Experimental Setup

3.1 Test Material

Recordings covering five Czech short vowels—/a/, /e/, /i/, /o/ and /u/(Czech SAMPA notation)—in all consonantal contexts were made in an anechoic room by two professional speakers—male and female. The recorded script was composed of three word sentences containing CVC word in the middle each, e.g. `/kra:lofski:` **kat** `konal/`. Recorded data were re-synthesized using the "half sentence" method [11]. This method consists in cutting the sentences in the middle of the vowels in the central words and combining the left and right parts, which results in a large set of sentences containing only one concatenation point in the middle of the central CVC word each and covering the vowels in all possible consonantal contexts. Note that the concatenations were done pitch synchronously using a simple overlap-and-add and weighting by the Hanning window, but no smoothing algorithm was applied for the reasons explained in [12].

3.2 Definition of Nasalization Mismatch

Since this work deals with the nasal contexts, a selection was made that only contained synthetic sentences containing a *nasalization mismatch*, henceforth referred to as the NAMI set (NAMI stands for the NAsalization MIsmatch). The rest of the sentences formed a set NOMI (NO MIsmatch). The *nasalization mismatch* was defined as a disagreement between an original context of a vowel and its target context, no matter if the disagreement was on the left or on the right side of the synthesized vowel. As an example, let us take two words `/t_San/` and `/t_Sas/`, and create a synthetic word `/t_Sa-as/` (dash symbol marks the concatenation point) using the left part of the first word and the right part of the second one. In our analysis, this synthetic word would be considered as containing the *nasalization mismatch*, because the original right context of the vowel /a/ in the first word was the nasal /n/, whereas in the synthetic word, the right context is the fricative /s/.

3.3 Mitigating the Role of F0 Discontinuities

F0 discontinuities are unquestionably a significant source of concatenation artifacts [12]. In order to be able to analyze the effect of nasalization, it was needed to factor the F0 discontinuities out. In most related studies, the standard procedure is to smooth the concatenation points with respect to differences in pitch and energy to make sure that any perceived discontinuity is not due to pitch or energy "jumps" at the concatenation points. As mentioned above, we have decided not to apply any pitch smoothing algorithm during concatenation, which was mainly to avoid the risk of introducing any sort of $F0$ smoothing artifacts that could influence listeners' ratings.

We have shown that clustering of pitch contours and concatenating words whose pitch contours fall within the same clusters is not a reliable way of predicting F0 concatenation artifacts [13]. Still, the information contained in the pitch contours can be leveraged for predicting concatenation discontinuities with a high accuracy using machine learning techniques [14].

For this work, we applied the SVM models trained on fine grained pitch contours extracted from a vicinity of concatenation points (see details in [14]) to identify sentences of the NAMI set that were supposed to be smooth, i.e. not containing an audible discontinuity at the concatenation points, according to the models' prediction. Let us further refer to this set as NAMI-S (NAsalization MIsmatch Smooth). The same models were analogically applied to obtain a set NOMI-S (NO MIsmatch Smooth).

3.4 Listening Test

Test Stimuli. As the next step, we randomly selected pairs of sentences—one sentence from the NAMI-S set and one from the NOMI-S set—containing the same word in the middle. Each vowel was represented by 15 pairs of sentences, resulting in the total number of 150 audio samples per voice presented to listeners. The sequence of pairs of samples of different vowels was also randomized. Two listening tests were organized—one for the male voice and one for the female voice.

Subjects. The subjects taking part in the listening tests were TTS experts and students working on TTS related projects. There were 9 and 10 subjects who finished the male and the female voice listening tests, respectively. Most of the subjects participated in both listening tests.

Procedure. The listeners were presented with the pairs of audio samples in a randomized order. Their task was to indicate whether or not they heard a concatenation discontinuity in any of the two samples, and which of the two samples they found better. It was also possible to say that none of the samples in a pair was better.

Both listening tests were conducted using a web interface allowing the listeners to work from home. It was, however, stressed in the test instructions that the tests shall be done in a silent environment and using headphones. As a preparation, the participants were presented (prior to each listening test) with a couple of samples containing audible discontinuities. There were no restrictions on how many times the listeners could play each sentence before providing their answers.

4 Listening Test Evaluation

Since the reliability of results obtained in any listening test—no matter if the participants are experts or not—is always an issue, a listeners ratings analysis was conducted in line with the procedures described in [15]. Based on the analysis results, one listener was excluded from each listening test.

As the next step, we have collected two sets of "facts". The first—discontinuity detection—set of "facts" was composed of audio samples for which more than or equal to 80% of listeners indicated that they heard a discontinuity (henceforth refferred to as

Table 1. *Counts of the discontinuity detection "facts" (DISC_FACT)*

	Female		Male	
	NAMI-S	**NOMI-S**	**NAMI-S**	**NOMI-S**
/a/	14	0	0	0
/e/	15	1	1	0
/i/	6	0	0	0
/o/	9	0	2	0
/u/	1	1	0	0

Table 2. *Counts of the preference "facts" (PREF_FACT). None stands for no preference "facts".*

	Female			Male		
	NAMI-S	**NOMI-S**	**None**	**NAMI-S**	**NOMI-S**	**None**
/a/	0	15	0	0	0	3
/e/	0	11	0	0	1	4
/i/	1	5	0	0	1	4
/o/	0	10	0	0	2	6
/u/	2	3	0	0	0	6

DISC_FACT). The second set of "facts" was based on preference scores, i.e. stimuli for which more than or equal to 80% of listeners expressed their preference for one of the samples in a pair or indicated no preference (henceforth referred to as PREF_FACT). Note that the majority threshold was set ad hoc to obtain a reasonable compromise between robustness and quantity of the "facts".

The results of the "facts" collection are summarized in Tab. 1 and Tab. 2. It is obvious from both tables that the *nasalization mismatches* at the concatenation points do not matter for the male voice, which was used in our study. This is in contrast to the female voice where we can see that especially for the vowels /a/, /e/ and /o/, there is a strong impact of the *nasalization mismatch* on the perceived quality of concatenations. To less extend, the effect can also be found for the vowel /i/.

It was interesting to speculate about the reasons for the obtained results from the perspective of the theory of the speech production. Since the problems were observed mainly in the female voice vowels /a/, /e/ and /o/, one interesting hypothesis that arose, was that the perceptual effects were due to a complex interaction of spectral peaks in a low frequency range.

For the female speakers' high vowels (/i/ and /u/), there are three spectral peaks in the frequency range between 200 and 500 Hz—fundamental frequency (F0), as well as the first oral and nasal formants (F1 and N1, respectively); see, for instance, [9]. It is well known that frequency components lying within 3 to 3.5 Bark from each other tend to be perceptually integrated into one broader peak [16]. That is exactly the case with the spectral peaks mentioned above.

It was therefore possible that the concatenation of an oral and a vowel from a nasal context could result in some sort of discontinuity acoustically, but since it is the energy within this 3.5-Bark band, which is relevant perceptually, the discontinuity could be inaudible. Supposing that this was true, it would have explained why /i/ and /u/ behave differently—the other vowels' F1 lies in higher frequencies and therefore falls outside of the 3.5-Bark range of the perceptual integration. It would have also explained why we did not find a similar situation for the male voice—his F0 lies much lower than the peak complex of F1 and N1.

5 Analysis of Discontinuities

To verify the hypothesis formulated in the previous section, we more closely investigated the concatenation areas in both time and frequency domains, and we got an intriguing finding. As shown in Fig. 1, the reason for the perceived discontinuities was a phase mismatch at the concatenation points. The phase mismatch appeared as a consequence of misplacement of pitch marks by our pitch marking algorithm [17], which got confused by strengthening of a harmonic signal component the peaks of which were in close vicinity of the F0 peaks. This strengthening of the harmonic component appears when a vowel (/a/, /e/ or /o/, to be more precise) stands in a context of a nasal consonant. For the vowel /a/, all audible discontinuities can be fixed by manual relabeling of the pitch marks in the concatenation areas, which was confirmed by an informal listening test.

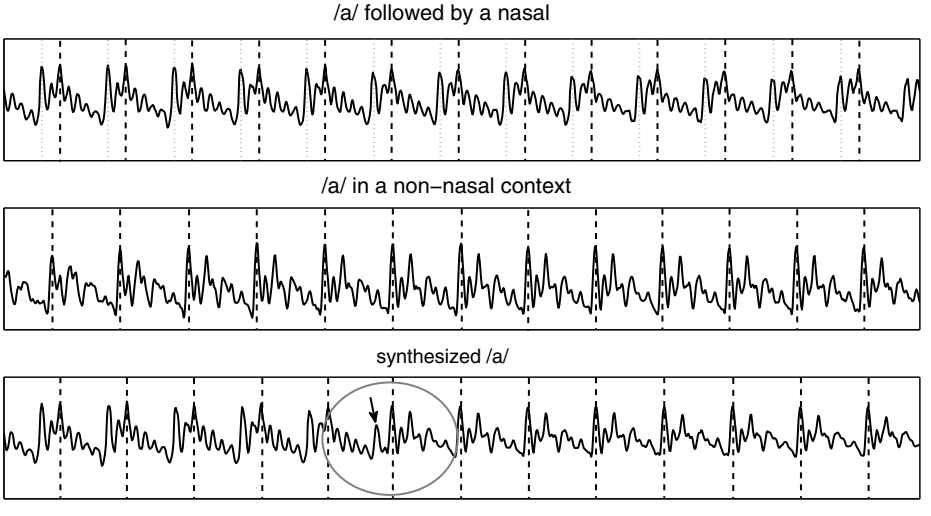

Fig. 1. Illustration of phase mismatch at a concatenation point due to mislabeling of pitch marks. Dashed lines show the positions of pitch marks as originally given by the automatic pitch marking algorithm. The dotted lines show the manual corrections of the pitch marks' positions. The circled area contains a pointer to an artificial signal peak due to the phase mismatch at the concatenation point.

The situation was however more complicated for the vowels /e/ and /o/, for which the perceived discontinuities could not be removed in this way in all sentences. The reason was that the harmonic component mentioned above got not only stronger, but also interfered with the F0 peaks. In such cases, it seems to be more advisable to avoid concatenations at all.

Interestingly enough, for the male voice we did not observe this phenomenon, which also explains why the *nasalization mismatch* does not matter for this particular voice.

Regarding the high vowels of the female voice, the discussed harmonic component is very weak in both nasal and non-nasal contexts. More important factor seemed to be energy differences at the concatenation points (especially for /i/). It appears to be, however, perceptually of less importance, as the listeners found the discontinuity "facts" in a smaller number of sentences.

6 Conclusion and Future Work

In this paper, we have closely investigated the effects of nasal context mismatches on the quality of concatenations in vowels for two Czech speakers—male and female. The results clearly showed that the *nasalization mismatches* have a strong effect on perceived quality of concatenations in vowels /a/, /e/ and /o/ for our female speaker. Upon closer inspection of spectrograms and oscilograms of the concatenation points, it was found out that the concatenation discontinuities were due to strengthening of a harmonic component of the vowels in nasalized contexts, which in most cases resulted in phase mismatches at the concatenation points. In some cases, when the harmonic component interferes with the F0 peaks, the concatenation artifacts can not be removed by fixing the phase at the concatenation points and it is therefore better to completely avoid such concatenations. For the male voice, no impact of the *nasalization mismatches* was found.

Since the study was conducted for two voices only, it needs to be repeated using various speakers, and possibly also in more languages, in order to draw some more general conclusions. Still, we believe that the findings shown in this paper are an important indication of what we have so far been missing when measuring quality of concatenations in the diphone based concatenative speech synthesis. We have made similar observations for different consonantal phonetic context mismatches, and we plan to publish the results of these analyses in near future.

References

1. Klabbers, E., Veldhuis, R.: Reducing audible spectral discontinuities. IEEE Transactions on Speech and Audio Processing 9, 39–51 (2001)
2. Bellegarda, J.R.: A novel discontinuity metric for unit selection text-to-speech synthesis. In: SSW5 2004, Pittsburgh, PA, USA, pp. 133–138 (2004)
3. Vepa, J.: Join cost for unit selection speech synthesis. Ph.D. thesis, University of Edinburgh (2004)
4. Syrdal, A.K.: Phonetic effects on listener detection of vowel concatenation. In: EUROSPEECH 2001, Aalborg, Denmark, pp. 979–982 (2001)

5. Syrdal, A.K., Conkie, A.: Perceptually-based data driven join costs: comparing join types. In: INTERSPEECH 2005, Lisbon, Portugal, pp. 2813–2816 (2005)
6. Kawai, H., Tsuzaki, M.: Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative speech synthesis. In: ICSLP 2002, pp. 2621–2624. Denver, Colorado (2002)
7. Fujimura, O., Lindqvist, J.: Sweep-tone measurements of vocal-tract characteristics. J. Acoust. Soc. Am. 49, 541–558 (1971)
8. Fant, G.: Acoustic theory of speech production. Mouton, The Hague (1960)
9. House, A.S., Stevens, K.N.: Analog studies of the nasalization of vowels. J. Speech Hearing Disorders 21, 218–232 (1956)
10. Hawkins, S., Stevens, K.N.: Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels. J. Acoust. Soc. Am. 77, 1560–1575 (1985)
11. Legát, M., Matoušek, J.: Design of the Test Stimuli for the Evaluation of Concatenation Cost Functions. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 339–346. Springer, Heidelberg (2009)
12. Legát, M., Matoušek, J.: Analysis of Data Collected in Listening Tests for the Purpose of Evaluation of Concatenation Cost Functions. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 33–40. Springer, Heidelberg (2011)
13. Legát, M., Matoušek, J.: Identifying Concatenation Discontinuities by Hierarchical Divisive Clustering of Pitch Contours. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 171–178. Springer, Heidelberg (2011)
14. Legát, M., Matoušek, J.: Pitch contours as predictors of audible concatenation artifacts. In: Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, pp. 525–529 (2011)
15. Legát, M., Matoušek, J.: Collection and Analysis of Data for Evaluation of Concatenation Cost Functions. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 345–352. Springer, Heidelberg (2010)
16. Chistovich, L.A.: Central auditory processing of peripheral vowel spectra. J. Acoust. Soc. Am. 77, 789–805 (1985)
17. Legát, M., Matoušek, J., Tihelka, D.: On the detection of pitch marks using a robust multi-phase algorithm. Speech Communication 53, 552–566 (2011)