



Temporal alignment of creaky voice in neutralised realisations of an underlying, post-nasal voicing contrast in German

Tina John, Jonathan Harrington

Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University, Munich, Germany

tj@phonetik.uni-muenchen.de, jmh@phonetik.uni-muenchen.de

Abstract

The aim of the present experiment was to investigate the acoustic phonetic cues that could underlie a post-stress voicing distinction which, when considered on a segmental basis, appears to be neutralised. The data concern the difference in German between minimal pairs such as 'Enten' and 'enden' which in more casual speaking styles appear to show schwa and oral stop deletion and a surface realisation as a neutralised creaky voice nasal. We extracted from the Kiel Corpus all such contrasts that were judged by trained transcribers to have been neutralised in this way. We measured the spectral slope over the first two harmonics and the time at which the spectral slope first changed significantly. Our results show that, contrary to segmentally-based assumptions, /t, d/ were distinguished depending on the onset time of creaky voice relative to the preceding vowel. These data are consistent with a model in which cues to segmental contrasts may be distributed non-segmentally in time.

1. Introduction

As is well known, there is a stop voicing contrast in German that is said to be neutralized in syllable-final position. The literature gives different approaches on how to represent this contrast phonologically as extensively discussed by [1]. In some cases the feature [spread glottis], defined articulatorily as an active glottal opening gesture, is used to model the stop voicing distinction, whereas in other models the distinction is based on [voice] i.e., on whether or not there is a quasi periodic opening and closing of the vocal folds. The feature [spread glottis] assumes a non 'spread glottis' for /b,d,g/ and a 'spread glottis' for /p,t,k/, whereas [voice] assumes a quasi periodic opening and closing of the vocal folds in a segment /b,d,g/ or non vocal fold vibration in /p,t,k/.

In a minimal pair in which stops are flanked by nasals - such as 'Enten' ('ducks') vs. 'enden' ('to end') which at an abstract phonological representation may be represented as /ɛntən/ vs. /ɛndən/, there is in a more casual speaking style no medial schwa and there may be no evidence of an oral stop so that the final sequence reduces to a long /nn/ sequence.

In [2] Kohler investigated glottalisation in the production of oral stops at all places of articulation (/b,d,g,p,t,k/) in a more casual speaking style. Instead of the oral stops Kohler observed a simple glottal valve action involving either a glottal stop or a more relaxed glottalisation. The creaky phonation is used to cut off the air stream to imitate the typical oral stop closure. In the flanking nasal context in which the oral closure may be at the same place of articulation as the nasal, it is common that the oral stops surface as a neutralised creaky voiced nasal. In the case of apical nasals, Kohler discusses different temporal alignments of creaky voice with the sonorant /nn/. In words like

'könn/t/en' ('could') or 'Stun/d/en' ('hours') he observed [ɲɲ] realisations for canonical underlying voiceless stops, [nɲ] for voiced stops and [nɲn], [ɲɲ] for voiced and voiceless stops. In a perception experiment, listeners were asked to identify stimuli with or without spliced glottalisation as an example of /nən/ or /ntən/. The presence of glottalisation was found to be a cue for an oral stop for listeners, while the precise synchronization with the nasal was ignored by the listeners [2], what might be caused by the experimental design. Kohler [2] neither discussed whether the different alignments of creaky voice are systematically due to the voicing distinction of the underlying stops nor whether these different alignments are perceptual cues to different voice qualities of the stop.

Gordon and Ladefoged describe creaky phonation acoustically as a "series of irregular spaced vocal pulses" [3:386] with lower power, lower F0 and lower spectral slope compared to modal phonation. According to [3,4], the spectral slope based on the first two harmonics in a short time power spectrum seems to be most reliable for discriminating modal and creaky phonation. For creaky voice, the spectral slope is positive while it is negative for modal and breathy phonation [3]. This is the case if the amplitudes of the harmonics are not boosted too much by the resonant frequencies of the oral tract. To avoid such an influence, the source spectrum that results from an inverse filtering [5] is more suitable for measurements of the spectral slope and thus for the detection and discrimination of creaky and modal phonation.

Although the occurrence of [ɲɲ] for apical nasal articulations is mentioned by Kohler [2], there are few details about how creaky voice is aligned with neighboring segments. Hawkins and Nguyen [6] have observed that stop voicing cues may be temporally remote from the site of the segmental contrast: specifically, they showed that voicing cues for voiced or voiceless syllable codas may extend as far forward as the onset of the same syllables. Their findings demonstrate that "words can be recognized from relatively weak auditory information spread across more than one acoustic-phonetic segment [...]" [6:225].

Local [7] claims the inter-relationships and temporal synchronization of phonetic parameters ("parametric interpretation") are important for interpreting the functionality of the speech signal. In contrast to the more common interpretation of segmentally-based features he claims that "features or set of features over different domains (e.g., phrasal units, words, syllables, syllable constituents)" [7:336] ("variable-domain interpretation") are equally important.

The present study follows the proposals of Local and the findings of Hawkins and Nguyen and investigates the acoustic phonetic cues that could underlie a post-stress voicing distinction. Therefore we reconsidered Kohler's investigation of oral stops in a complete nasal environment (see above). Most interesting are the [ɲɲ] realisations mentioned by Kohler

[2], where glottalisation seems to be temporally aligned to the nasal only. In contrast to Kohler, we do not limit our observation to the creaky voiced nasal but extend it to the preceding segment. For this reason we examined the extent and temporal alignment of creaky voice using a measure of spectral slope over the first two harmonics in suitable utterances from the Kiel Corpus of Read and Spontaneous speech [8]. Our data include canonical underlying /ndən/ and /ntən/ sequences in non utterance final words that were judged by trained transcribers to have been neutralised by /ə/ elisions and in which the oral stops were replaced by creaky voice. Further exponents of underlying /nən/ were included as control sequences because these are produced with a modal voice over the whole duration of the sequence.

2. Method

2.1. Materials

The materials included /nn/ realisations of underlying /ntən/ and /ndən/ with /ə/ and stop elisions and transcribed with creaky voice (e.g., [n̥n], [n̥n̥], [n̥n̥]). We also included the control sequence /nən/ in which /ə/ elision had also taken place. All such sequences were extracted from the Kiel Corpus of Read (27 male, 26 female speakers of Standard German) and Spontaneous (24 male and 18 female speakers of Standard German) Speech together with the preceding vowel in non utterance final words. The /nən/ data were only extracted from the Kiel Spontaneous Speech corpus.

Table 1 gives the absolute number (n) of the occurrence of the different sequences analysed in this study. The data include at least one utterance of each vowel in the German vowel system except [u:] and [aʊ].

Table 1. *The total number of tokens (n) per category analysed in this study.*

Sequence	n.	Example
Vntən	244	könnten, fünfzehnten
Vndən	71	Stunden, jemanden
Vnən	55	keinen, Terminen

2.2. Procedure

We analysed the temporal interval between the onset of the vowel and offset of the final nasal (henceforth [Vnn]) for all tokens.

The segments were divided into 20 equal time slices over the [Vnn] interval and f0 data and short time spectra with a 40 Hz resolution were calculated for each time slice (Fig. 1).

Each spectrum was inverse filtered using the following methodology in order to obtain a closer approximation to the glottal source. Firstly, we calculated the cepstrum [9] by applying a DCT transformation [10] to the signal's spectrum; secondly, we filtered out the cepstral coefficients 2 to 28 thereby leaving information in the cepstrum predominantly due to the glottal source (which occurs at higher cepstral coefficients) and to the combined mean and slope of the original spectrum (the lowest two coefficients). We then applied a DFT to this filtered cepstrum to obtain an approximation to the source spectrum.

For each estimated source spectrum, the spectral slope (m) over the frequency range (f_{low} - f_{high}) of the fundamental and 2nd harmonic (H1,H2) were calculated. To ensure that both harmonics were within the interval and to allow a 20%

margin of error in the f0 calculation f_{low} and f_{high} were calculated from a pitch-synchronous pitch-tracker by:

$$f_{low} = 0.8 f_0 \quad (1)$$

$$f_{high} = 2.2 f_0 \quad (2)$$

That way we collected 20 spectral slope values across the duration of the segment [Vnn] for each token of each sequence.

Finally, the slope, m_t at proportional time point t between the segment onset and offset was defined as the slope of the linear regression equation that was calculated by a least square estimation [11], where the frequency values were the predictors and the dB values at the frequency the responses.

We calculated 20 slopes (m_t) per segment at 20 equally spaced proportional time points ($t = 5, 10, 15, \dots, 95, 100\%$ of the segment's total duration).

We then applied a t-test with Bonferroni correction in order to determine whether m_t differed between the three underlying categories /ndən/, /ntən/ and /nən/ (we did this at each time point i.e., at $t = 5, 10, \dots, 100\%$).

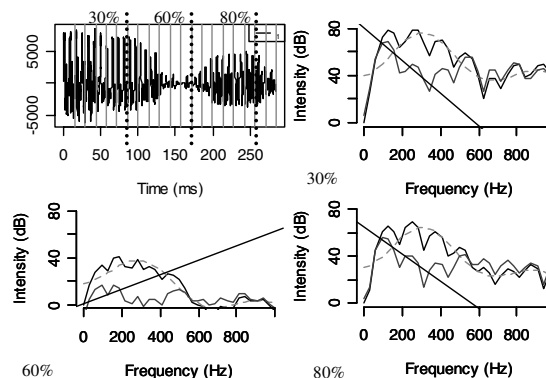


Figure 1. *Example of inverse filtering and slope (m_t) calculation with $t=30\%$, 60% , 80% of segment duration for one token of /ndən/ with original spectrum (black), filter spectrum (grey dotted), glottal spectrum (grey).*

The duration of the vowel (V) and the nasal (nn) as well as the duration of the [Vnn] dyads were measured for each of /ntən/, /ndən/ and /nən/. t-tests were applied to test for significant differences between the three categories.

3. Results

3.1. Preliminary investigation

The duration measurements in which we compared the durations of the three types under investigation in [Vnn] showed the following results. Firstly, we found significantly longer vowel durations ([Vnn]) for /nən/ than for /ntən/ ($t = -2.920$; $df = 83$; $p < 0.001$) and /ndən/ ($t = -3.903$; $df = 106$; $p < 0.001$) while the latter two did not differ from each other. Secondly, for the nasal ([Vnn]) significantly shorter durations were found for /nən/ than for both /ntən/ ($t = 6.917$; $df = 71$; $p < 0.01$) and /ndən/ ($t = 5.384$; $df = 107$; $p < 0.001$). /ntən/ and /ndən/ durations again did not differ. Thirdly, the duration of the entire vowel+nasal duration ([Vnn]) did not differ between the three categories.

We had initially restricted our analysis to the differences between /ntən/ vs. /ndən/ vs. /nən/. However, it soon became apparent that we were not able to identify accurately the onset time of creaky voice – and for this reason, we extended the analysis to include the preceding V in all three cases.

3.2. /nən/ slopes

t-tests applied to successive proportional time points in /nən/ showed the only difference to be between the very onset and offset of the segment. Moreover, as Figure 2 shows, there is scarcely any change in the median slope throughout the segment. From this we conclude that a predominantly modal voice was used throughout /nən/.

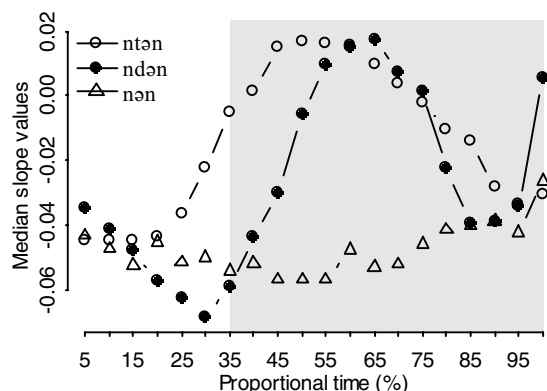


Figure 2: Median values of all spectral slopes marked at 5% intervals of the total segment duration for /Nn(t)dən/. The grey box represents approximately the results of the preliminary investigation i.e. from the offset of the vowel to the offset of the sequences.

3.3. /ndən/ and /ntən/ slopes

It is evident from Figure 2 that there is a change in the spectral slope in /ndən/ and /ntən/ during the sequence of vowel and nasal. Near the onset of these sequences, the slopes are negative and their variation is minimal. For both sequences of interest an increase into positive values and then a decrease down to negative values is observable in the middle to the end of [Vnn]. Although there is a similar movement in the slopes for both sequences, they differ in their temporal alignment. The main point of distinction is where the slopes change from negative to positive. For /ntən/ this is at $t = 40\%$ and for /ndən/ at roughly $t = 55\%$. Thus, the /ntən/ slopes rise earlier or faster. The t-tests with Bonferroni correction showed no significant differences between /ndən/, /ntən/ and /nən/ up to $t = 20\%$. At $t = 25\%$, there is a significant difference between /ntən/ and /nən/ ($p < 0.001$) and between /ntən/ and /ndən/ ($p < 0.001$). $t = 50\%$ is the first time point at which /ndən/ differs significantly ($p < 0.001$) from /nən/ and at this time point there is no significant difference on slopes between /ndən/ and /ntən/ anymore.

Beyond $t = 55\%$, the slopes of /ntən/ and /ndən/ decrease and there is no significant difference between them until the last time point. At $t=85\%$ for /ndən/ and $t=90\%$ for /ntən/ until the end, the slopes of these sequences are both equal ($p > 0.01$) to those of /nən/.

4. Discussion

Following [6], creaky phonation produces a positive spectral slope in the H1-H2 frequency range, whereas it is negative for modal phonation. The results on the /nən/ slopes gave no evidence of creaky voice in the realisations of /nən/ because there are no positive slope values. There is, by contrast, evidence of creaky voice in /ntən/ and /ndən/ (because some of the slopes are positive).

The main finding of this investigation is that when /ntən/ and /ndən/ are reduced to a long /nn/ sequence, there are nevertheless differences between them in the timing of the onset of creaky voice. Specifically creaky voice starts earlier in the voiceless sequence. This result is robust across several speakers, speaking styles, and contexts.

With this measure, we were able to establish that voice quality changes from modal to creaky voice and back to modal phonation over the /V(t)dən/ interval. We were also able to show that up to 20% of the /V(t)dən/ duration, /ndən/ and /ntən/ were similar to /nən/ and thus all produced with modal phonation. We then established that creaky voice in /ntən/ extends over a proportionally wider time interval compared with /ndən/. Towards the segment offset, /ndən/ and /ntən/ were indistinguishable and both produced with modal phonation.

The preliminary investigation of the slopes without the preceding vowel showed the same tendency but the modal phonation for /ntən/ at the onset of [Vnn] was missing. Thus extending the study to the adjacent vowel is justified and moreover was found to be necessary to detect the time at which the modal phonation started turning into creaky phonation.

It is clear that a segmentally-based analysis is inadequate because we cannot say that the onset and offset of creak is restricted to any segmental boundaries. Moreover, a segmentally-based approach misses the fine phonetic details of creaky voice synchronization that we have found here. Following Local [7] and Hawkins and Nguyen [6], we believe it is likely although not yet demonstrated that this type of fine phonetic detail may be important in the perceptual distinction of the stops in these contexts.

A uniform feature representation for the voiced/voiceless stop contrast across all possible contexts is also insufficient: our results show that the features [voice] or [spread glottis] are insufficient for distinguishing between /t/ and /d/ in this context. Voiced and voiceless oral stops or their occurrence in a nasal context and the kind of spontaneous or casual speaking style that characterized the stimuli of the present investigation are produced with adducted (i.e. weakly vibrating) vocal folds. But the features [spread glottis] and [voice] define a contrast between adducted and adducted vocal folds which is clearly not appropriate for the stimuli examined here. The phonetic differences between the voiced and voiceless stops examined here depend on different temporal alignments of creaky voice which cannot be easily expressed by static features such as [spread glottis] and [voice].

5. Conclusions

The current investigation has shown that there is a different temporal alignment and extent of creaky voice phonation in [nn] realisations of underlying /ntən/ and /ndən/. In /Vntən/, creaky voice is aligned with the vowel preceding the nasal whereas creaky voice for /Vndən/ begins during the following

nasal. We conclude that a feature representation of the stop voicing contrast in German cannot be modeled by a feature contrast assigned segmentally but needs to include fine phonetic details that are temporally remote from the site of the contrast.

6. Acknowledgements

This research was supported by German Research Council grant HA 3512/2-1 to Jonathan Harrington and Christine Mooshammer.

7. References

- [1] Jessen, M. and Ringen, C., "Laryngeal features in German", *Phonology*, vol. 19, pp. 189-218, 2002.
- [2] Kohler, K., "Investigating unscripted speech: implications for phonetics and phonology", *Phonetica*, vol. 57, pp. 85-94, 2000.
- [3] Gordon, M., Ladefoged, P., "Phonation types: a cross-linguistic overview", *Journal of Phonetics*, vol. 29, pp. 383-406, 2001.
- [4] Hanson, H. M., Stevens, K., Kuo, H. J., Chen, M., Slifka, J., "Towards models of phonation", *Journal of Phonetics*, vol. 29, pp. 451-480, 2001.
- [5] Ní Chasaide, A., Gobl, C., "Voice source variation", In: Hardcastle, W.J., Laver, J. (Eds.), *The Handbook of Phonetic Sciences*, Oxford: Blackwell, pp. 427-461, 1997.
- [6] Hawkins, S., & Nguyen, N., "Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English", *Journal of Phonetics*, vol. 32, pp. 199-231, 2004.
- [7] Local, J., "Variable domains and variable relevance: interpreting phonetic exponents". *Journal of Phonetics*, vol. 31, pp. 321-339, 2003
- [8] Simpson, A. P., Kohler, K. J., Rettstadt, T. (Eds.), *The Kiel Corpus of Read/Spontaneous Speech - Acoustic data base, processing tools and analysis results*. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), vol. 32, 1997.
- [9] Childers, D.G., Skinner, D.P., Kemerait, R.C., "The Cepstrum: A Guide to Processing", *Proc. of the IEEE*, 65, 1977.
- [10] Abdi, H., "(The Method of) Least Squares", In. Salkind, N.J. (Ed.), *Encyclopedia of Measurement and Statistics*, Thousand Oaks (CA), Sage, pp. 530-532., 2007.
- [11] Milner, B. and Shao, X., "Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end", *Speech Communication*, vol. 48, pp. 697-715, 2006.