Simulating hypernasality with phonological features in Swedish TTS

Alexander Näslund[1], Christina Tånnander[2,3], Sofia Strömbergsson[4], Marcin Włodarczak[1]
[1]Department of Linguistics, Stockholm University, Sweden
[2]Swedish agency for Accessible Media
[3]Speech, Music and Hearing, KTH, Sweden
[4]Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Sweden
alexander.naslund@ling.su.se, christina.tannander@mtm.se, sofia.strombergsson@ki.se, wlodarczak@ling.su.se

Abstract

This study explores the use of phonological features in text-to-speech (TTS) synthesis by simulating hypernasal resonance. By training neural TTS with phonological features instead of more traditional input methods, we are capable of producing speech sounds unseen in the training data, letting us manipulate the voice in our current task: to produce disordered speech. The results of an acoustic analysis indicate successful manipulation of hypernasal resonance, although a perceptual evaluation showed lacking authenticity, particularly in certain prosodic aspects. Despite limitations, our approach using synthetic speech with simulated hyper-nasal resonance shows potential in the use of perceptual training for speech-language pathology (SLP) students, provided that relevant improvements are implemented.

Introduction

Hypernasal resonance, or *hypernasality*, is a resonance disorder typically caused by speech motor difficulties or structural deficits in the speech apparatus. Hypernasality manifests as nasal resonance on speech sounds where none is expected, making vowels nasalized while also affecting voiced consonants. As a symptom, it is known to be difficult to assess successfully for licensed SLP practitioners. Reliable assessment of hypernasality is therefore key for ensuring suitable clinical care but learning material for perceptual assessment training is not always available.

Text-to-speech (TTS) synthesis has gone through a rapid development for the past decade in the production of synthetic speech, achieving human-like results (see e.g. Tan et al., 2021). However, its potential in simulating disordered speech is not nearly as explored.

Using phonological features as input to TTS gives us the ability to map features of the voice to a range of numerical values, which can then be manipulated by increasing or decreasing the value. This method allows TTS synthesis to produce speech sounds which are absent from the training data. For example, nasality could be increased for all speech sounds in a sentence, creating nasal resonances where there otherwise would not be any.

In this study, we aim to explore whether the method used can generate speech with the acoustical characteristics that is inherent in hyper-nasality. The generated speech will be measured and analysed both acoustically and perceptually.

Background

Phonological features and TTS

With "Preliminaries to Speech Analysis", Jakobson et al. (1952) defined speech sounds as a set of features based on articulatory criteria. These features were binary, for example Vocalic/Consonantal, Grave/Acute and Nasal/Oral. Their work is viewed as part of the foundation for what is called phonological features today.

As input method for neural TTS synthesis, phonological features are more detailed and less commonly used, but gives us increased control over the output compared to the more common TTS input methods, such as graphemes or phonemes.

In our approach, these features are represented by gradual values between 0 and 1 for various degrees of a specific feature. The task of the TTS model is then to learn the relations between the phonological feature values and the speech in the training data. During synthesis, it is possible to use feature values that did not occur in the training data as input, including values below 0 and above 1. This makes it possible to synthesize speech sounds which the original speaker never produced.

Little research has been done on the topic of simulating disordered speech, and even less combined with the use of phonological features as input. Moëll et al. (2022), used phonemic feature vectors to synthesize speech data augmentation for automatic phoneme recognition of aphasic speech (phonemic paraphasia). Moëll et al. used a set of binary features proposed by Chomsky and Halle (1968). The resulting TTS voice was used as training material to improve automatic speech recognition (ASR) models for patients with aphasia, showing improved results. Another study used phonemes as input to TTS, simulating dysarthric speech also to serve as training data for ASR with noticeable improvements to the ASR model (Soleymanpour et al., 2022).

Hypernasality

Hypernasality is defined as a resonance disorder caused by a velopharyngeal dysfunction. The dysfunction results in insufficient closure of the velopharyngeal port, creating nasal resonance also during the production of oral speech sounds. In addition, the insufficient closure allows air leaking through the nasal cavity during the production of plosives, causing voiced plosives to turn into voiced, nasal speech sounds.

Acoustically, nasalised vowels are characterised by the existence of additional resonances and anti-resonances in the vowel caused by the coupling between the nasal and the oral cavity. The nasal resonances will enhance harmonics relative to their surroundings which results in “nasal peaks” or “nasal formants” (Styler, 2017a). Also, due to “nasal zeros” or “nasal antiformants” occurring in the nasal cavity, the voice should see a reduction of spectral energy in certain frequency ranges.

The most recommended method for measuring nasal resonance in vowels is A1-P0 (Styler, 2017a) in which we subtract P0, a nasal formant occurring between 250 to 450 Hz, from A1, the amplitude of F1. The resulting value is lower the more nasal resonance exists in a vowel due to A1 being lowered by the nasal zeros and P0 rising due to the nasal resonances.

Given that A1-P0 is affected by specific vowels, Chen (1996) recommends A1-P0 to be adjusted for vowel types and offers a function which uses the relative bandwidths and position of formants. This “compensated” measurement will be referred to as A1-P0(comp) throughout this paper.

Method

The TTS voice

The model used in this study was trained on over 12 000 Swedish sentences read by a female professional speaker, originally recorded for unit selection TTS by the Swedish Agency for Accessible Media, as part of the Filibuster project (Ericsson et al., 2007).

The data was originally segmented into phonemes, and each phoneme were then converted into a phonological feature vector of 15 phonological features and one language feature. Each feature was assigned a value between 0.0 and 1.0. Consequently, the nasality feature was 0.0 for oral speech sounds and 1.0 for nasal consonants and vowels. Oral vowels in nasal contexts were given +0.25 nasality for each adjacent nasal speech sound.

The neural TTS voice used was trained using OverFlow (Mehta et al., 2023) as encoder and HiFi-GAN (Kong et al., 2020) as vocoder. Overflow is based on a neural hidden Markov model (HMM) with normalizing flows. The decoder/vocoder HiFi-GAN is a Generative Adversarial Network (GAN) designed for high-fidelity speech synthesis. HiFi-GAN performed faster and produced more human-like speech than the best publicly available models of its time during its release in 2020 and Overflow was shown to produce speech with fewer mispronunciations than comparable methods in 2023.

Test data

For the acoustic analysis of vowels, the five peripheral vowels /iː ɛː uː oː ɑː/ were chosen. The vowels were put between voiced alveolar or bilabial plosives as context, either /b/ or /d/ as both onset and coda and inserted in the carrier phrase “Jag säger --- igen”, (English: “I'm saying --- again”). Five degrees of nasality were applied, ranging from 0.0 to 1.0 with a 0.25 increment, affecting all speech sounds by the same value for all sentences. Recordings from the human voice underlying the training of the TTS voice were included as reference in the analysis, referred to as “test data” throughout this paper.

The sentences used for the perception test were taken from SVANTE (Lohmander et al., 2015). These sentences are designed to enable assessment of articulation and nasal resonance, through their consonant patterns and lack of nasals. Five sentences were chosen:

1. Bibbi bara jobbar
2. David å du leder
3. Kicki kokar korv
4. Giggi vill väga guld
5. Lollo lurar Ella

The values used for the degrees of simulated nasality differed from the data used in the acoustic measurements. In this case, the values 0.33, 0.67 and 1.0 were used for all speech sounds, as well as 0.67 and 1.0 for vowels only, resulting in a total of five different nasality settings. The reason for the split between affecting all speech sounds and affecting vowels only was to investigate which of these two methods provided a more perceptually accurate simulation of hypernasality.

A1-P0

A1-P0 were measured using Styler's and Scarborough's (2017b) Praat (Boersma & Weenink, 2024) script for nasality measurement. The script required the audio files to be annotated in Praat Textgrids with two tiers; vowel and word. Annotation of the words were done manually by finding the beginning and end of the target word in each carrier phrase. Annotation of the vowels were done by identifying a 0.05 second interval in the middle of the vowel. The script was set to run with standard settings except for the number of measurements per vowel, set to three. This setting means that the script measured the vowel at three separate points, in the beginning, middle and end of the annotated segment.

Consonant substitution

A manual annotation was made on the same words used in measuring A1-P0(comp) but only for the degrees of 0.5, 0.75 and 1.0 because of it being unlikely that any substitutions would take place if

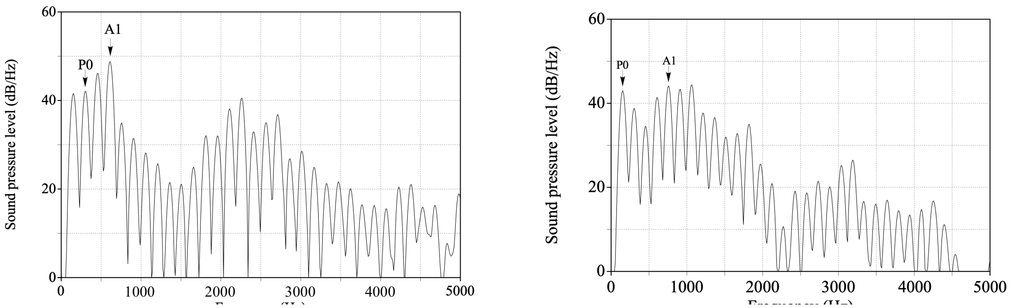

Figure 1. Spectral slices for synthesized /ɛː/ with the applied nasality degrees of 0.0 (left) and 1.0 (right).

no, or close to no, manipulation takes place (degree 0.0 and 0.25). The annotation was done by the first author who counted the plosives and nasals, respectively, in the target word. If a plosive had been substituted by a nasal it was also documented if the place of articulation remained the same or if it changed. Each word had two plosives each, which made the data used into a total of 40 words and 80 consonants per nasality value.

Perception test

Five licensed speech pathologists, the majority of whom were some of the leading experts on hypernasality in Sweden, took a survey made in Google Forms. They were instructed to use headphones and to assess the degree of hypernasal resonance in manipulated speech, with the alternatives None, Slight, Moderate and Severe. Three additional questions concerned their experience of hypernasality assessment, how they experienced the speech samples in the perception test, and finally if they found any potential in using simulated speech disorders in perceptual skill training for SLP students.

Results

Figure 1 shows spectral slices for the nasality degrees 0.0 and 1.0 in the manipulated voice. An increase in P0 and a decrease in A1 is indeed visible when nasality is set to 1.0, indicating that the manipulation of nasality worked as expected. In addition, the degree of 1.0 was associated with other acoustic markers of nasality, such as presence of antiformants, indicated by the sudden drop in energy around 500 Hz and 2300 Hz. We can also see an increase in energy around 1000 Hz, most likely corresponding to P1, another nasal formant.

Figure 2 shows A1-P0(comp) for all the data points separated by the different degrees of applied nasality in the simulated voice together with the recordings made by the original speaker (test data). As can be seen there is a decrease in A1-P0 for the highest degree of manipulation (1.0) indicating an increased nasal resonance in vowels. The differences show no obvious pattern in the lower values. Notably, the recordings from the test data deviates slightly from the unmanipulated voice (0.0) by showing a slightly wider spread of data points.

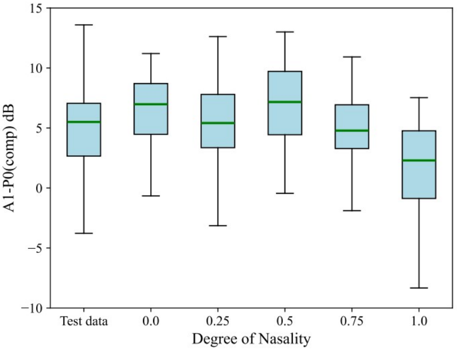

Figure 2. A1-P0(comp) for all vowels

The results from the manual annotation of consonants with nasality degrees 0.5, 0.75 and 1.0 can be seen in Table 1. It shows a clear shift in nasal substitution the more simulated nasal resonance is applied to the voice, with almost every plosive substituted by a nasal when the nasality was set to 1.0.

Table 1. Number of substitutions for every word and value split by place of articulation. The values range from 0 to 8, where 8 is the maximum number of substitutions possible for a given word and degree.

Word	0.5	0.75	1.0	Place
/biːb/	0	0	0	bilabial
	0	0	8	alveolar
/bɛːb/	0	0	7	bilabial
	0	0	0	alveolar
/bɑːb/	0	3	8	bilabial
	0	0	0	alveolar
/boːb/	0	0	0	bilabial
	0	0	8	alveolar
/buːb/	0	0	0	bilabial
	0	0	8	alveolar
/diːd/	0	0	0	bilabial
	0	0	8	alveolar
/dɛːd/	0	0	0	bilabial
	0	0	7	alveolar
/dɑːd/	0	0	0	bilabial
	0	8	8	alveolar
/doːd/	0	0	0	bilabial
	0	4	8	alveolar
/duːd/	0	0	0	bilabial
	0	3	8	alveolar

Some differences between the vowels did occur. At 0.75 almost all the nasal substitutions happened in /ɑː/ for both contexts with some occurrences in /uː/ and /oː/ as well. At 1.0, all the plosives had turned nasals except one of the occurrences of /bɛːb/.

The plosives substituted by nasals generally retained their place of articulation, though not always: the bilabial plosives in /biːb/, /boːb/ and /buːb/ were substituted to the more posterior position by the alveolar /n/.

For the bilabial contexts, /bɑːb/ and /bɛːb/ were the only words where the plosives kept their place of articulation. The alveolar contexts followed the same pattern, where the plosives were replaced by the alveolar nasal /n/.

The perception test involved both quantitative and qualitative questions, of which the quantitative results can be seen in Table 2. Here, the median score was taken for the combined answers for each audio file, grouped by degree of nasality. 1.0 on all speech sounds received a median score of "Moderate" while the lower degrees all received a median score of "None".

Table 2. Median scores from the perception test for each degree of nasality applied on all speech sounds or vowels only.

Degree of nasality	Applied on	Median
0.33	all	none
0.67	vowels	none
0.67	all	none
1.0	vowels	none
1.0	all	moderate

The respondents agreed that simulated speech disorders could potentially be used in perceptual assessment training for SLP students, but there was also a consensus that the provided audio files were not good enough for this purpose. Comments concerned the quality of consonants being deviant in an unrealistic manner and the intonation was increasingly deviant the higher nasality degree was applied. It was also noted that real-life patients rarely exhibit hypernasality and no other symptoms. Lastly, a few remarks were made about the lack of credibility due to the speaker not being a child.

Discussion

This study sought to find out whether the manipulated voice would show the acoustical characteristics that is inherent in hyper-nasality. This hypothesis proved to be correct, indicated by the facts that A1-P0 were lower at the higher degrees of manipulation and that all consonants except one out of 80 were substituted by a nasal for the highest degree of manipulation (1.0).

A1-P0 is proven to work well for measuring nasal resonances in vowels (Styler, 2017a). One interesting finding of the present study was that the TTS voice showed a generally lower median score of A1-P0 and a slightly larger spread of data points than the recordings of the speaker's natural voice used as training data. Given that it is technically a different voice with different contexts used for the vowels, both of which are factors proven to influence A1-P0, this finding is in line with previous research (Styler, 2017a).

Regarding nasal substitution, the involvement of only one annotator is a limitation, preventing evaluation of annotation reliability. If the substitution would have been between speech sounds more alike each other, a more precise way of measurement would have been appropriate.

It could have been interesting to include more natural test data sentences, as a complement to the standardized SVANTE sentences and the carrier phrases with target words, both for the acoustic measurements and in the perception tests.

The open questions in the survey provided us with important information and gave us further insights in the potential of the TTS voice. There was a consensus among the participants that simulated speech disorders of this kind could be beneficial for perceptual skill training in speech therapy studies, provided that certain aspects of the synthesized speech were improved to enhance authenticity.

The TTS voice used was trained using some of the best performing models currently available. Regarding the features used, we used a set of phonological features that is currently under development. It is designed for experimentation with non-binary features in Swedish neural TTS. In this study, the only feature that was manipulated and used in a non-binary manner was the investigated nasality feature. Using a fully binary features set would have excluded the possibility of investigating different degrees of simulated nasality. The rest of the feature set matches a standard Swedish phoneme description and is in practice equivalent to typical phoneme-based training.

Conclusions

The TTS voice managed to create synthetic speech with hypernasal resonance, demonstrated by a reduced A1-P0 in vowels and substituted/reduced consonants. We conclude that the synthesized speech with simulated hypernasal resonance shows some potential, but is not yet suitable for direct application in SLP e-learning or educational settings.

Overall, this study contributes to advancing the understanding of using phonological features in TTS synthesis. Future research could focus on fine-tuning the feature manipulation, combining commonly co-occurring speech disorders commonly found together as well as exploring using a child TTS voice.

Acknowledgements

We would like to thank the individuals who participated in the survey for their valuable answers.

References

Boersma, P., & Weenink, D. (2024). Praat: Doing phonetics by computer [computer program]. version 6.4.06. http://www.praat.org/

Chen, M. Y. (1996). Acoustic correlates of nasality in speech. Acoustic Correlates of Nasality in Speech," Ph. D. dissertation, Massachusetts Institute of Technology, Cambridge, MA

Chomsky, N., & Halle, M. (1968). The sound pattern of English. Harper Row, New York

Ericsson, C., Klein, J., Sjölander, K., & Sönnebo, L. (2007). Filibuster: A new Swedish text-to-speech system. Fonetik 2007, Stockholm, 50(1), 33–36

Jakobson, R., Fant, C. G. M., & Halle, M. (1952). Preliminaries to speech analysis. The distinctive features and their correlates. Cambridge, MA: MIT Press

Kong, J., Kim, J., & Bae, J. (2020). Hifi-Gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33, 17022–17033 https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

Lohmander, A., Borell, E., Henningsson, G., Havstam, C., Lundeborg, I., & Persson, C. (2015). SVANTE - Svenskt Artikulations- och Nasalitetstest. Manual (2. uppl). Studentlitteratur AB.

Mehta, S., Kirkland, A., Lameris, H., Beskow, J., Székely, É., & Henter, G. E. (2023). OverFlow: Putting flows on top of neural transducers for better TTS. Proc. INTERSPEECH 2023, 4279–4283 https://doi.org/10.21437/Interspeech.2023-1996

Moëll, B., O'Regan, J., Mehta, S., Kirkland, A., Lameris, H., Gustafson, J., & Beskow, J. (2022). Speech data augmentation for improving phoneme transcriptions of aphasic speech using Wav2Vec 2.0 for the PSST challenge. Proceedings of the RaPID Workshop, 62–70 https://aclanthology.org/2022.rapid-1.8/

Soleymanpour, M., Johnson, M. T., Soleymanpour, R., & Berry, J. (2022). Synthesizing dysarthric speech using multi-speaker TTS for dysarthric speech recognition. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7382–7386. https://doi.org/10.1109/ICASSP43922.2022.9746585

Styler, W. (2017a). On the acoustical features of vowel nasality in English and French. The Journal of the Acoustical Society of America, 142(4), 2469–2482 https://doi.org/10.1121/1.5008854

Styler, W., & Scarborough, R. (2017b). Nasality automeasure v. 5.9 https://github.com/stylerw/styler_praat_scripts/blob/master/nasality_automeasure/NasalityAutomeasure.praat

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. https://doi.org/10.48550/arXiv.2106.15561