# Perceptual Consequences Of Nasal Surrogates In English: Implications For Speech Synthesis[1]

Susan R. Hertz[§*], Isaac Spencer[§*], Richard Goldhor[§], & Tom Church[*]      [§]NovaSpeech LLC & [*]Cornell University      www.novaspeech.com

## ABSTRACT

Previous experiments have demonstrated that non-nasal obstruents in human utterances can be replaced by a wide range of "surrogate" segments, either produced by formant synthesis or recorded from other speakers, with virtually no change in perceived speech quality or speaker identity [1]. The experiment summarized in this poster extends the previous work by exploring the perceptual effects of nasal obstruent segment substitution. The results suggest that (1) vowels adjacent to nasal surrogates must have appropriate acoustic features to cue the perceptually desired nasals, (2) the nasal surrogates themselves must have generally appropriate prosodic characteristics for their context, but (3) specific spectral properties of nasal surrogates are not perceptually important for cuing speaker identity, segmental identity, or overall naturalness. It appears that natural sounding speech quality and the speaker's identity can be preserved even when a speaker's nasal consonants are replaced by surrogate segments possessing strikingly different acoustic structure than the originals', including nasals from different segmental contexts, nasals of different phonemic identity, nasals from different speakers (including different genders), formant synthesized nasal segments, and non-nasal segments.

## INTRODUCTION

This experiment uses segment splicing techniques to explore three hypotheses about syllable-initial nasals in continuous speech in English:

| Hypotheses about syllable-initial nasals in continuous speech | Predictions |
|---|---|
| In most contexts, the internal spectral structure of syllable-initial nasals is perceptually irrelevant for their segmental identity. | A wide variety of spectrally diverse surrogate speech segments can be substituted in place of the original segments with minimal perceptual consequence. |
| Syllable-initial nasals do not contain perceptually important information about a specific speaker's voice. | Surrogates from different speakers can be employed without affecting judgments of speaker identity. |
| Nasalization in vowels adjacent to syllable-initial nasals can be a strong cue to the perception of those consonants as nasals, as well as to specific speaker identity. | Vowel surrogates must have contextually-appropriate nasalization for proper speaker and nasal consonant perception. |

## METHODOLOGY

We elicited acceptability, intelligibility, naturalness, and speaker identity judgments on a variety of renditions of the sentence *Monica Naimoo never knew Bonnie's mother* ("the Monica sentence") in which various phoneme-sized segment substitutions were made. This sentence was selected for several reasons: it contains the two possible syllable-initial English nasal phonemes in a range of phonological and phonetic contexts; it contains proper names that can not be deduced from context; it contains vowels that contrast in their degree of nasalization, but are otherwise similar for our speakers in the two contexts; it contains consonant-vowel sequences in which the formant transitions have desired characteristics (e.g., in which formants transitions move in the same direction out of both [n] and [m]); and other reasons.

We recorded and digitized the speech of three speakers: two middle-aged males (JS and JD) and a middle-aged female (SH). We constructed stimuli by replacing nasal consonants or their adjacent vowels in the JS and SH Monica sentences with a variety of different segment types, as shown below:

| Types of Segment Substitutions Tested | |
|---|---|
| Nasal consonants from one speaker for those of another | Oral consonants in place of nasal consonants |
| Nasal consonants from one segmental context into another | A synthetic consonant in place of a natural consonant |
| A single nasal phoneme in place of all nasal phonemes | F0-compatible and F0-incompatible nasal consonants in place of nasal consonants |
| Silence in place of nasals | |
| Oral vowels in place of nasal vowels | |

Listeners were told that they would be judging stimuli consisting of the Monica sentence and related sentences that had been processed in different ways for a particular purpose, such as a speech encoding system or special type of telephone system. Elicited responses would include judgments concerning acceptability, naturalness, segmental intelligibility, speaker identity, and signal quality.

A NovaSpeech program called Nisper (NovaSpeech Integrated Speech Perception Experiment Runner) was used to present stimuli at a listener-controlled pace, and collect and store the listeners' responses. The sample screenshot below illustrates the Nisper user interface:

**Figure 1: Sample Nisper User Interface**

Listeners were introduced to the experimental tasks in a training session during which they became familiar with natural renditions of the Monica sentence as spoken by SH and JS. (One of their tasks would involve judging the extent to which certain stimuli sounded like these speakers.)

For each stimulus, listeners were required to choose one of the five possible quality categories shown in the following table:

| Overall Quality Categories | |
|---|---|
| **Rating** | **Meaning** |
| Extremely Natural | Stimulus sounds like an unprocessed, fully-natural speech utterance. |
| Very Natural | Stimulus sounds highly natural, but some processing effects are noticeable. |
| Acceptable | Stimulus sounds reasonably natural and human-like. It is clearly intelligible, but may exhibit noticeable artifacts resulting from speech processing, such as static or choppiness. |
| Poor | Stimulus exhibits obvious problems, like strong background noise or buzziness. |
| Very Unnatural | Stimulus quality is totally unacceptable: listener would be unwilling to listen to speech that sounded like this. |

For any stimulus receiving a score below Very Natural, listeners were also required to mark one of the ten specific problem categories shown in the Nisper display in Figure 1. The specific problem categories were intended to provide insight into the questions shown below:

| The problem checkboxes were designed to answer questions like these: |
|---|
| Did the sentence sound human? |
| Did the sentence sound like it was spoken by either SH or JS? |
| Did the sentence sound like it was spoken by a mix of speakers? |
| Did the sentence sound like the speaker had a cold? |
| Did the sentence have any phonemes that differed from those in the Monica sentence? |
| Did the sentence contain extraneous non-speech glitches or noise? |

Nineteen listeners each judged 26 sentences drawn from a cohort of 38 stimuli. The speech stimuli were presented through loudspeakers. Each stimulus was presented twice per session in a fixed arbitrary order. Listeners could listen to each stimulus as many times as desired. For each stimulus, listeners first had to play and judge the entire stimulus. Following this overall judgment, they marked problems with specific subsections of the utterance, as shown in Figure 1, but they could not change their Overall Quality rating. If listeners perceived a phoneme difference between the stimulus and the Monica sentence, they were instructed to specify which segment was different and, if possible, what phoneme they perceived.

## STIMULI

Stimuli were constructed by hand through waveform concatenation using the Praat speech analysis tool [2]. Surrogate segment waveforms were spliced into the base utterance by abutting them without overlap. All speech was digitized at 22050 Hz. with the exception of two 8000 Hz stimuli that were used to test the effect of sampling rate on perceptual judgments. With a few exceptions, the duration of each surrogate was adjusted to match that of the segment being replaced. The F0 of most phonetically voiced surrogates was interpolated between the preceding and following segments using Praat's PSOLA-based F0 smoothing facility [3]. For experimental purposes, the original F0 of some segments was left unaltered, to explore the perceptual consequences of F0 incongruity.

Figure 2 to the right shows spectrograms of the first two words, *Monica Naimoo*, of the Monica sentence as spoken by the three source speakers, SH, JS, and JD. Note the range of prosodic and spectral variation among the nasal consonants across contexts and speakers.

Table 1 to the right lists the source utterances from which all of the stimuli were constructed.

Table 2 below shows the composition of each stimulus. The color of each stimulus fragment indicates the source utterance from which the fragment originated. The first half of the table depicts the stimuli that use SH as the base utterance, the second half JS.

Along the top of Table 2 are the phonemes of the Monica sentence. Within each stimulus, each surrogate appears in the same column as the phoneme it replaces. The surrogate is indicated by a letter giving its original phonemic identity.

If our hypotheses predict that the surrogate will not alter the perception of the original phoneme, an unbolded font is used; if they predict a different phoneme will be perceived, a **bold underlined font** is used. A "?" is used for original (non-replaced) consonants whose perception we thought might be affected by the substitution of a neighboring vowel (e.g. oral for nasal).

A white "X" indicates that no F0 smoothing was used. Figure 3 shows portions of two stimuli (4 and 5) that differ only in whether F0 smoothing was used.
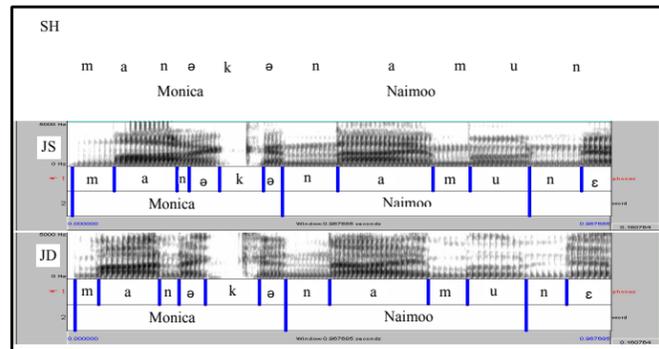
**Figure 2: Each speaker's rendition of *Monica Naimoo* . . .**

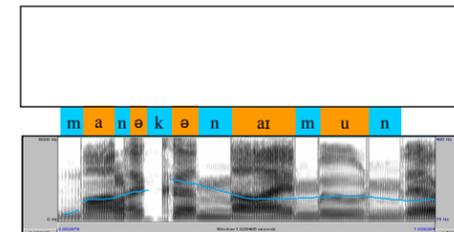| Utterance ("U") # | Speaker | Utterance | Stimuli used as Orig. | Stimuli used as Sur. |
|---|---|---|---|---|
| 1 | SH | *Monica Naimoo never knew Bonnie's mother.* | 1-13 | 3, 18 |
| 2 | JD | | | 6, 19 |
| 3 | JS | | 15-26 | 4-5, 17 |
| 4 | SH | *Bodica Daiboo deber due Bodie's brother.* | | 11-13 |
| 5 | JS | | | 24-26 |
| 6 | SH | *Monica Naimoo never knew Mommy's mother.* | 14 | |
| 7 | JS | *Monica Maimoo never knew Bonnie's mother.* | 27 | |
| 8 | SH | *Tonica Naitoo never knew Donnie's brother.* | | 8, 10 |
| 9 | JS | | | 22 |
| 10 | synthetic | synthesized [n] | | 7 |
| 11 | SH | [ama] (spoken slowly) | | 9 |
| 12 | JD | [ana] (spoken slowly) | | 20-21 |
| 13 | SH | [asa] (spoken slowly) | | 10 |

**Table 1: Stimulus Sources**

**Figure 3: Nasal surrogates with discontinuous vs continuous F0 (fragments of stimuli 4 and 5)**

The blue line overlaid on the spectrograms is a pitch trace produced by the Praat pitch tracker.

**Table 2: Stimulus Composition. (Colors Refer to Source Utterance in Table 2.)**

As can be seen, the data strongly support our prediction that spectrally diverse surrogate speech segments, even from other speakers, can be substituted for nasal consonants without altering those consonants' phonemic identity. There were 4,090 responses to substitutions for which we predicted that the original phoneme would be perceived, and, as shown in column 3, in virtually all these cases, the original phoneme was indeed perceived.

### Speaker Identity

The following table gives statistics for three general substitution types on how often the intended speaker (that of the base utterance) was misidentified.

| Speaker Identity Results | | | |
|---|---|---|---|
| Substitution Type | # of Responses | % ID Change "Other" | % ID Change "Mult" |
| Nas Cons Substitution, Same Speaker | 891 | 0% | 3% |
| Nas Cons Substitution, Diff Speaker | 2,356 | 3% | 7% |
| Oral-for-Nasal Vowels, Same Speaker | 76 | 33% | 22% |

The data support our prediction that surrogates from other speakers can be employed without altering speaker identity. Of the three substitution types, only the oral-for-nasal vowels, which we did not expect to sound natural, were frequently marked as sounding like a speaker other than SH or JS ("Other") or as sounding like a combination of different speakers ("Mult").

It is worth noting that four listeners knew SH and JS well, but none of them detected any "Other" speaker surrogates in the stimuli. In fact, two of the listeners were the speakers' own children!

The "Mult" speaker judgments generally resulted from discontinuous F0 patterns of the sort illustrated in Figure 3 (stimuli 5 and 21 in Table 2).

## SELECTED RESULTS

This section highlights some of the main results concerning acceptability, segmental identity, and speaker identity judgments.

### Acceptability

The following table summarizes overall acceptability in terms of two measures: (1) average overall quality ("OQ") scores, where 5 = very natural and 1 = very unnatural, and (2) the average number of problems reported for a stimulus (see Nisper user interface in Figure 1).

| Overall Stimulus Acceptability | | | |
|---|---|---|---|
| Substitution Type | Acceptability Measures Sample Count | Avg. OQ Scores | Avg. # Problems reported |
| Original Consonant Surrogates (stimuli 3 and 17) | 52 | 3.6 | 1.5 |
| Consonant Surrogates, Same Speaker | 50 | 3.3 | 1.9 |
| Consonant Surrogates, Other Speaker | 238 | 3.3 | 1.8 |
| Oral-for-Nasal Vowels, Same Speaker | 152 | 2.4 | 2.3 |
| F0 Incompatibility | 68 | 2.3 | 2.3 |

As a reference for the overall quality judgments, the table shows the scores for stimuli 3 and 17, in which the original nasal consonants were extracted and spliced back in using the same splicing technique used for the stimuli in general. These stimuli had natural voice quality, but somewhat degraded signal quality as a result of some splicing artifacts. It turned out that listeners were highly sensitive to such artifacts in their acceptability judgments, a fact that we attribute to the nature of the instructions given to the listeners and the fact that listeners could play utterances as many times as they wished before responding.

Note that the reference stimuli received only slightly better overall quality measurements (3.6) than the stimuli in which nasals were replaced by a wide range of surrogate types (3.3). Using surrogates from other speakers did not significantly affect either the overall speech quality or the number of problems reported.

In fact, the only stimuli considered unacceptable were those in which nasalized vowels were replaced by oral surrogates, and those in which substituted nasal consonants had incompatible F0.

Very few users reported that any of the stimuli were unintelligible (data not shown), and with few exceptions, only those stimuli with incompatible F0 were marked as sounding non-human.

### Segmental Identity

The following table summarizes the extent to which our predictions concerning segmental identity were borne out by listeners' responses.

| Segmental Identity Results | | |
|---|---|---|
| | # of Responses / % Correct Prediction | Prediction |
| Substitution Type | Substituted phonemes will be perceived | Original nasal will be perceived |
| Same Speaker, Same Phoneme | 54 / 82% | 360 / 100% |
| Same Speaker, Different Phoneme | 186 / 87% | 1,980 / 99% |
| Different Speaker, Same Phoneme | No samples | 1,444 / 100% |
| Different Speaker, Different Phoneme | No samples | 306 / 100% |

## CONCLUSION

This experiment is part of a larger project in which we are investigating a new unified theory of speech perception and its application to speech synthesis. Previously we had noted that non-nasal obstruents are cued not so much by their internal spectral structure, but rather by a combination of their duration, their amplitude relative to neighboring segments, and the acoustic structure of neighboring vowels. The syllable-initial nasal consonants tested in this experiment yielded similar results: as long as a nasal's F0 is congruous with surrounding vowels, a wide range of acoustically-diverse consonants can serve as nasal surrogates, including segments from different speakers, contexts, and phonemes. The results have a number of implications for speech synthesis. Both rule and unit selection strategies can be simplified by ignoring the perceptually irrelevant details in speech, such as nasal murmurs.

## REFERENCES

[1] Hertz, 2002: Integration of rule-based formant synthesis and waveform concatenation: a hybrid approach to text-to-speech Synthesis: *Proc. IEEE 2002 Workshop on Speech Synthesis*.

[2] Praat: Boersma, P. and D. Weenink, Univ. of Amsterdam, www.fon.hum.uva.nl/praat (as of May 23, 2004).

[3] Psola: E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453-467, Dec 1990.

[4] Klatt, D. H. and Klatt, L. C. (1990). Analysis, Synthesis and Perception of Voice Quality Variations among Female and Male Talkers, *J. Acoust. Soc. America* **87**, 820-857.

## ACKNOWLEDGMENTS

## NOTES

[1]This is a revised version of a poster presented at the 147th Meeting of the Acoustical Society of America in New York, May 2004.

For a more general overview of our perceptual theories and related work see http://www.novaspeech.com/projects/mit_2004.pdf .