Effects of Source-Tract Interaction in Perception of Nasality

António Teixeira\textsuperscript{\dagger}, Francisco Vaz\textsuperscript{\dagger} and José Carlos Príncipe\textsuperscript{\ddagger}

\textsuperscript{\dagger}Dep. Electrónica e Telecomunicações/ INESC Aveiro, Universidade de Aveiro
3810 Aveiro, Portugal email: ajst@inesca.pt

\textsuperscript{\ddagger}Computational NeuroEngineering Lab, University of Florida, Gainesville, USA

ABSTRACT

In this paper we study the effect of source changes, caused by vocal tract load, in perception of nasality. For that we have developed an articulatory speech synthesizer, including a comprehensive nasal tract model and an interactive glottal source model. Our main objective was to investigate to what extent is necessary, in systems aimed to produce high quality synthetic sounds, to include the effect of source-tract interaction in the glottal source model when synthesizing nasal vowels. In our studies we used Portuguese nasal vowels. Portuguese uses nasalization of vowels in its phonological inventory. Changes in glottal wave, caused by the additional load of the nasal tract are more significant in vowels like [\textipa{\textcircled{a}}] with low $F_1$ and high $F_2$. Effects are more dramatic in time rather frequency domain. Perception tests favor the idea that listener aren’t able to detect the perceptual effect of source-tract interaction changes caused by the additional coupling of the nasal tract. More tests are needed to support, or reject, this.

1. Introduction

It is well known that the shape of the glottal excitation is of primary importance to the produced quality of synthetic speech. Several studies have addressed this problem but nasal sounds were not included in these studies. Since effects of nasalization occur in the first formant region, also reported in the interaction studies as the principal responsible for the effect of tract load, it is our opinion that effects of nasalization in the source deserve investigation. Also we know that nasal sounds have more attenuated high frequencies. This can be motivated, in part, by source spectral properties.

To study the effect of interaction we started by doing some simulations that addressed the change in tract input impedance, glottal wave and spectral properties of the glottal wave.

Two perception tests were also performed. The first test aimed to detect if listeners are able to detect differences in the synthetic nasal vowels with and without the inclusion of nasal tract load to obtain the glottal excitation. The second test aimed to study which of the stimuli were preferred. The studies used synthetic stimuli produced by an articulatory synthesizer that enabled to generate nasal sounds with dynamic variation of articulators positions.

2. The Articulatory Synthesizer

2.1. Overall Description

The overall description of the synthesizer has been reported in [10]. Briefly, it consists of an articulatory model, representing the vocal tract in the sagittal plane, and an acoustic model including both oral and nasal tracts. The first gives the cross-sectional area along the vocal tract and the second simulates the propagation in the tube. There is also a glottal source model. For the present work, the nasal and interactive source model are of primary importance.

2.2. Comprehensive Nasal Tract Model

From the characteristics of the nasal cavities and of the Portuguese nasal sounds we arrived at the following requirements for the model: (1) capable of work with occlusions of the nasal tract; (2) it should be possible to use symmetric and asymmetric models; (3) possibility of inclusion of paranasal sinuses; (4) definition of the tract dimensions should be easy. All this characteristics make the synthesizer more comprehensive than the already existent. Implementation detail can be found in [11].

It is easy to use different nasal tract configurations in our synthesizer. The nasal configuration, including the sinus, is defined in a simple ASCII file. The method of calculation of the tract impedance used by the interactive source model, makes possible to the researcher to choose between including or not the nasal tract input impedance in the calculation of the tract impedance seen from the glottis.

2.2.1. Nasal tract dimensions

In this study, we used the nasal tract dimensions from [3] which were based on studies by Dang et al. [4] and Stevens [9].

One brief comment is due about the choice of the radiation area ($\text{Area=0.5 cm}^2$). The use of such a low radiation area was motivated by the need to match the acoustic properties of the pharyngonasal tract. A pharyngonasal configuration, using data from [7], has $F_1=300 \text{ Hz}$, $F_2=1000 \text{ Hz}$ and $F_3=1900 \text{ Hz}$. The paranasal sinuses are not enough to obtain such a low $F_1$. In [5] the use of a small radiation area is proposed to obtain such a low $F_1$. 
2.3. Interactive Glottal Source

It is well known that the "naturalness" of synthetic speech is closely related to the shape of the glottal pulse [8]. We do not yet have a complete understanding of the phonatory behavior of the vocal folds. Thus, we lack an efficient model of the voice source. However, several models capable of describing the major characteristics of the glottal flow have been proposed. They can be classified in two major categories: interactive and non-interactive models. In the interactive models, there are two approaches to generate the glottal volume velocity. In non-physical approaches, glottal flow is calculated by modeling the glottal area [2, 1] or conductance function and by incorporating the various impedances of the acoustic system into the model. For the physical approach, structural modeling of the mechanical vibration of the vocal cords [6] or kinematic model for the 3-d glottis has been attempted.

We implemented an interactive source model based in [1]. The lack of detailed information needed by physical models, added to the possibility of direct control, in non-physical model, of wave characteristics, such as fundamental frequency, explains our choice. Several improvements were made: inclusion of a two-mass parametric model of the glottal area; inclusion of irregularities such as Jitter, Shimmer and aspiration; and the capacity of synthesis of sounds with time varying source and/or tract parameters.

3. Simulations

We made several simulations to study the effect of nasality. The coupling of the nasal tract alters the transfer function, the input impedance of the tract. The change in input impedance alters load condition of glottis changing the glottal wave.

3.1. Effect in the Tract Input Impedance

We looked at effects of nasalization in tract input impedance. In Figure 1 we compare the input impedance of an oral vowel, an [a], with the nasal vowel with the same oral tract configuration. Also shown is the impedance of an nasal consonant produced by closing the lips, maintaining all the other articulators positions. Input impedance shows no profound change, but there are a few alterations. There are extra peaks, especially in the $F_1 - F_2$ region. The nasal vowel case is almost indistinguishable from the nasal consonant case.

3.2. Effects in the Glottal Wave

Glottal wave changes due to nasal tract coupling are rather small. Simulations for vowels [˜a] and [˜ı] are presented in Figures 2 and 3. First figure compares the oral to the nasal vowel (with and without oral closing). Again, lip closing effect is very small. Second figure presents also no tract load condition for comparison purposes. For the nasal vowel 2 cases are presented. One using the complete tract input impedance, the other not including nasal tract load. $u_g(t)$ in both cases is different. Nasal tract load seems to influence the wave. This change is more noticeable in vowels [˜ı] and [˜e], with low $F_1$ and high $F_2$. More studies are necessary.

Figure 1: Input impedance for a oral vowel, nasal vowel and a bilabial nasal consonant formed by closing the lips without any other articulators movement.

Figure 2: Glottal wave and respective Fourier transform of the vowel [a]. 3 cases are presented: (a) Velum closed, (b) open velum, (c) open velum and closed lips- configuration of an bilabial nasal consonant. All cases using total source tract interaction.

3.2. Effects in the Glottal Wave

Glottal wave changes due to nasal tract coupling are rather small. Simulations for vowels [˜a] and [˜ı] are presented in Figures 2 and 3. First figure compares the oral to the nasal vowel (with and without oral closing). Again, lip closing effect is very small. Second figure presents also no tract load condition for comparison purposes. For the nasal vowel 2 cases are presented. One using the complete tract input impedance, the other not including nasal tract load. $u_g(t)$ in both cases is different. Nasal tract load seems to influence the wave. This change is more noticeable in vowels [˜ı] and [˜e], with low $F_1$ and high $F_2$. More studies are necessary.
Much more simulations could be performed. Other vowels, simulations with other nasal tract configurations, different source parameters (e.g. open quotient as seen in Figure 4), could be simulated.

4. PERCEPTUAL TESTS

As shown in the previous section, the alteration of input impedance of the vocal tract, by the coupling of the nasal tract, changes the characteristics of the glottal wave. We want to know if these changes are perceived by the listener.

Perception studies using standard tests were conducted using synthetic stimuli produced by the articulatory synthesizer. General characteristics of used stimuli follow. Oral tract configurations were obtained by inversion using an optimization process. Resulting configurations were checked against bibliography reported configurations. Vulum aperture was obtained manually by adjusting the ratio of nasal coupling area to oral area, in velum region, to a value of around 10.

Timing, for vulum, was not very precise because we have no detailed production data. Information in the literature about velum transition times and analysis of natural vowels in CVC contexts was used. Same timing was used for all vowels. In the first 100 msec vulum stays closed, making an opening transition in 60 msec to the maximum value. Vulum remains at this maximum to the end of vulum. Stimuli end with a nasal consonant, a bilabial ([m]), produced by closing the lips. Closing movement of the lips starts at 200 msec ending 50 msec later.

Stimulus duration was fixed at 300 msec for all vowels. Regarding the source, the interactive source model was used with variable $F_0$. $F_0$ starts around 100 Hz raises 20 Hz in the first 100 msec and then gradually goes to 100 Hz. Open quotient was 60 % and speed quotient 2 [1]. Jitter and shimmer were added to improve naturalness.

To conduct the tests a computer program was developed to make the tests completely automatic. Program makes the stimulus pairs, repetitions, and randomization, presents stimuli to subject and records test results for further analysis. Signals were presented over Sennheiser Headmax HD 470 headphones in rooms with low ambient noise. Subjects responded by pressing, using the mouse, a button in the computer display.

A total of 9, 6 male and 3 female, European Portuguese native speakers participated in the test. Ages varied from 13 to 53 with mean 26.8.

4.1. Discrimination Tests

Are listeners able to perceive changes in the glottal excitation caused by the additional coupling of the nasal tract? A 4IAX discrimination test was performed to try to answer this question.

4.1.1. Stimuli

Stimuli were produced for 3 nasal vowels, [ã], [I] and [ʊ]. Only 3 vowels were used to reduce test realization time. Also, the other 2 Portuguese nasal vowels usually appear as diphthongs. Only dynamic velum stimuli were used because in previous work we found this stimuli as more natural [12]. The only factor varied was the input impedance of the tract used by the interactive source model that has 2 values: (1) input impedance includes the effect of all supraglotal cavities; or (2) the nasal tract input impedance is discarded.
### Table 1: Results of 4IAX test.

<table>
<thead>
<tr>
<th>Listener</th>
<th>Sex</th>
<th>[a]</th>
<th>[i]</th>
<th>[u]</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 EDU</td>
<td>M</td>
<td>50.0</td>
<td>33.3</td>
<td>41.7</td>
<td>41.7</td>
</tr>
<tr>
<td>2 CSST</td>
<td>M</td>
<td>58.3</td>
<td>100.0</td>
<td>50.0</td>
<td>69.4</td>
</tr>
<tr>
<td>3 ENE</td>
<td>F</td>
<td>50.0</td>
<td>41.7</td>
<td>50.0</td>
<td>47.2</td>
</tr>
<tr>
<td>4 MJT</td>
<td>F</td>
<td>33.3</td>
<td>83.3</td>
<td>66.7</td>
<td>61.0</td>
</tr>
<tr>
<td>5 FVAZ</td>
<td>M</td>
<td>16.7</td>
<td>58.3</td>
<td>33.3</td>
<td>36.1</td>
</tr>
<tr>
<td>6 AJST</td>
<td>M</td>
<td>66.7</td>
<td>66.7</td>
<td>66.7</td>
<td>66.7</td>
</tr>
<tr>
<td>7 VIT</td>
<td>M</td>
<td>50.0</td>
<td>50.0</td>
<td>41.7</td>
<td>47.2</td>
</tr>
<tr>
<td>8 RIC</td>
<td>M</td>
<td>58.3</td>
<td>58.3</td>
<td>41.7</td>
<td>52.8</td>
</tr>
<tr>
<td>9 PAT</td>
<td>F</td>
<td>41.7</td>
<td>50.0</td>
<td>66.7</td>
<td>52.7</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>47.1</td>
<td>60.1</td>
<td>50.9</td>
<td>52.7</td>
</tr>
<tr>
<td>std</td>
<td></td>
<td>15.2</td>
<td>20.8</td>
<td>12.9</td>
<td>11.3</td>
</tr>
</tbody>
</table>

4.1.2. **Procedure**

Each of the 4 combinations (ABAA, ABBB, AAAB, BBAB) was presented 3 times in random order. With this arrangement each pair to be tested appears 12 times. Order was different for each listener. Inter Stimuli Interval was 400 msec and Inter Pairs Interval was 700 msec.

4.1.3. **Results**

Table 1 presents the percentage of correct answer for the 4IAX test. From the table it is clear that listeners correct answers are not above chance. In fact t-tests, for each vowel and average over the 3 vowels, having by alternative hypothesis averages above chance (> 50%), were all non-significant at 5% significance level. Note that vowel [i] was correctly recognized more times.

4.2. **Preference Test**

Because we are interested in quality, we decide for the use of a quality test: the AB test. Despite the demand for more decisions by each listener, augmenting test duration, the paired comparison test is precise. Despite the results of the 4IAX we decide to perform this test. All vowels were tested this time.

4.2.1. **Stimuli**

Stimuli were produced for the 5 nasal vowels. For each vowel 1 degree of velum aperture and 3 types of tract input impedance (Total load, Oral tract only, no load). A total of 15 (5 vowels x 3 impedances) stimuli were used.

4.2.2. **Procedure**

The question asked to subjects was "Which of the 2 stimuli do you prefer as a Portuguese nasal vowel?" Listeners had four possible answers: "FIRST", "SECOND", "BOTH", and "NONE". Stimuli were presented 5 times both in AB and BA order. Inter Stimuli Interval was 600 msec.

4.2.3. **Results**

Only two comments about the results. Listeners had difficulty in choosing. In more than half of the pairs consistency rates were below 60%. Clearly stimuli are very similar perceptually. Analysis of consistent responses showed no clear trend in the preference of listeners. Some listeners prefer stimuli not including tract load, others total tract load. Several factors can contribute to this results. Glottal area is already skewed. Also nasal sounds can sound more clear not including the tract load. Listeners judged them better by being more clear, but better nasal sounds are not more clear.

### 5. Discussion

Simulations showed some, small, effects of the nasal tract load in the glottal wave time and frequency properties. Results of perceptual tests, conducted to study to what extent these alterations are perceived by listeners, support the idea that these changes are hardly perceptible. These results are in concordance with results reported in [13]. In their work Titze and Story reported that "An open nasal port...showed no measurable effect on oscillation threshold pressure or glottal flow".

Due to the many factors involved we can only consider the present work as a first in a series of experiments. Many parameters capable of influencing the results should be investigated, e.g. vowel duration, fundamental frequency, nasal coupling area, subglottal system. Also many improvements can be made to the models and parameters used in the articulatory synthesizer. Clearly tract configuration, for the nasal vowels, should be obtained by direct measures (e.g. MRI). Also we need production data regarding the glottal wave. We need direct measures of source parameters, such as fundamental frequency and open quotient, during Portuguese nasal vowels production. Concluding, we don’t have a final answer to the question, further work is needed.

### 6. Acknowledgments

Work was possible by the Ph.D. scholarship BD/3495/94 and project PRAXIS/P/PLP/11 222/1998.

### References


