INFLUENCE OF DYNAMICS IN THE PERCEIVED NATURALNESS OF PORTUGUESE NASAL VOWELS

António Teixeira*, Francisco Vaz*, and José Carlos Príncipe§

*Dep. Electrónica e Telec., Universidade de Aveiro, Portugal
§Computational NeuroEngineering Lab, University of Florida, Gainesville, USA

ABSTRACT

Several studies point to the need of regarding speech as a dynamic phenomenon. The influence of dynamic information in oral vowel perception is a subject of study for many years. In addition, some researchers point to the need to see nasal vowels as dynamic. To produce high quality synthetic nasal vowels, one of our interests that motivated this work, would be useful to know in what measure we need to include dynamic information. Our study consisted in a perceptual test using stimuli generated by an articulatory synthesizer. Our null hypothesis was that with the sound characteristics held constant, by maintaining velum and oral articulators fixed, quality is not statistically different from cases where velum and other articulators are allowed to move. Our results are against this null hypothesis pointing to the need of dynamic information to obtain nasal vowels of high quality.

1. INTRODUCTION

Quality of nasal vowels produced by current synthesizers needs improvement. One of the ways to contribute to the solution of this problem is to improve our knowledge about how they are produced and perceived.

1.1 Portuguese nasal vowels

Nasal vowels are one of the distinct characteristics of Portuguese because of their number and of their particular properties. According to a number of phonetic studies, the nasal vowels of Portuguese differ from the nasal vowels of, say, French, in that they are strongly nasalized near the end only. Some studies claim that Portuguese nasal vowels display nasality contours, that is, their initial half is oral; their final half is nasal.

A recent study of Brazilian Portuguese [15] reported that nasal vowels have three distinct phases: an oral start, followed by transition where the nasality is superimposed, ending in a nasal murmur. This nasal murmur is reported as being co-articulated with the vowel and not with the following consonant as reported in previous works. However, as referred in the study only experiments with synthetic stimulus can help confirm, or not, this hypothesis.

Listeners judgement of naturalness of nasal vowels are language dependent. Stevens, Andrade and Viana [17] reported that, when rating stimuli goodness or naturalness, French, Portuguese and English listeners preferred different amounts of temporal patterns. Portuguese prefer some nasal murmur and seem to have a slight preference for more nasalization in the vowel, than English, although not necessarily over the whole vowel.

1.2 Nasal vowel as dynamic

Feng and Castelli [4], based in simulation studies of French nasal vowels, presented "...the conception of the nasal vowel as a trend beginning with a simple configuration (the oral one), which is terminated in the same manner (the pharyngonasal one...). Nasalization for them can be "...considered as a dynamic trend from an oral configuration toward the pharyngonasal configuration."[Pp. 3704, 4]. In [p. 234, 10] is referred that "in addition to static cues, there are dynamic cues to nasal segments. Nasal consonants offer abrupt and therefore auditorily salient changes in the overall amplitude and spectrum of the acoustic signal...There may also be dynamic cues to distinctively nasal vowels. For example, M. Ohala found that in Hindi after an initial nasal consonant, a distinctively nasal vowel shoed progressively more velic opening during its production than did a comparable nondistinctively nasal vowel."

Several studies with oral vowels addressed the importance of dynamic information in perception (see [18] and [12] for a review). Study of dynamic information use in perception of nonoral vowels should also be done.

In our effort to produce high quality nasal vowels, using and articulatory model, we found that vowels produced with the appropriate velum aperture where perceived as nasal but not of natural quality. When we made the velum variable producing a nasal vowel consisting of the three phases, reported as characteristics of Portuguese nasal vowels, quality improved. These first studies where informal and with one vowel only. These preliminary studies, and theories, motivated our interest in testing the influence of velum dynamics in the quality of Portuguese nasal vowels. Our study addresses the question: Is the dynamics of the velum movement important in the production of high quality Portuguese nasal vowels?

Many of the reported studies concerning Portuguese nasal vowels have been performed using formant synthesizers. For other languages, like French, several studies have used articulatory synthesizers [5, 7, and 8]. We decided to try the use of this more complex tool. The first problem to address was the lack of an available synthesizer, which lead us to develop one with the necessary characteristics.

The first part of the present paper describes, briefly, the synthesizer developed to conduct this study. In the second part, we describe the experiment performed.

2. ARTICULATORY SYNTHESIZER

To conduct the studies we decide to use an articulatory
synthesizer to produce stimuli for perceptual tests. To make possible the study our synthesizer [19], based in [14], was improved by the addition of: (1) a comprehensive nasal tract model; and (2) an interactive source model.

![Articulatory Synthesizer main components](image)

Figure 1 – Articulatory Synthesizer main components.

2.1 General description
Briefly, the articulatory synthesizer contains three major blocks (Fig. 1): a source model; an articulatory model; and an acoustic model. The synthesizer is implemented in C/C++ and Tcl/Tk in a Linux PC system with a standard soundboard making the system inexpensive.

The sagittal articulatory model is an improved version of Mermelstein model [6]. Model articulators are tongue body, tongue tip, jaw, lips, velum and hyoid. Conversion from sagittal distances to area function uses empirical formulas proposed by Mermelstein [9]. In our acoustic model, we make the usual approximations when dealing with vowels: (1) propagation is planar; (2) the tract is straight; and finally the tube is approximated by the concatenation of elementary acoustic tubes of constant area. An equivalent circuit, represented by a transmission matrix, represents each one of these elementary tubes. Analysis of the circuit is performed in the frequency domain as in [14]. This technique makes the simulation of the frequency dependent losses possible. Simulation of losses is a major concern when dealing with the nasal tract. Synthetic sound is obtained by the convolution of the glottal wave with the impulse response of the tract, obtained by Inverse Fourier Transform of the transfer function [14]. The use of the convolution method avoids the problem of continuity of resonance in the faster method proposed by Lin. The use of a fast implementation of the IFIT minimizes the calculation time problem of the convolution method.

2.2 Modeling of the nasal tract
Because we are interested in the nasal sounds, the model of the nasal tract should be one major concern. It should be possible to define different nasal tract shapes and the inclusion of the several parasal sinus.

In general a single tube, motivated by simplicity concerns approximates the nasal tract. Nevertheless, recently, it was shown, using MRI data, the necessity of inclusion in the model of the two asymmetric lateral passages [3]. Models considering the two parallel passages should be possible as well more simplified models with a common tube.

Due to health problems, introduction of objects or deliberate blocking with the hands, the nasal tract can be totally of partially obstructed. Even with complete blocking, nasal vowels can be produced, and in this case, they sound different. It is important the possibility of blocking the nasal passages at any point.

2.2.1. Model description. The nasal cavity is modeled in a similar way to the oral tract. The major difference results from the fact that the area function of the nasal tract is fixed for the major part, for a particular speaker. The variable region, the soft palate, changes with the degree of nasal coupling. The velum parameter of the articulatory model controls this coupling. RLC shunt circuits representing Helmholtz resonators simulate the paranasal sinuses [7]. Our synthesizer permits the definition of different tract shapes and the inclusion of the needed sinus at any position by simply editing an ASCII file. Blocking can be done at any position by defining a null area section at the point of occlusion. Implementation details were reported in [20].

2.2.2. Nasal tract dimensions. In this study, we use the nasal tract dimensions from [2] which were based on studies by Dang et al. [3] and Stevens [16].

One brief comment is due about the choice, or more precisely of the nostrils radiation area used (Area=0.5 cm²). 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 [10], has F1=300 Hz, F2=1000Hz and F3=1900 Hz. The paranasal sinuses are not enough to obtain such a low F1. In [4] the use of a small radiation area (they used 0.6 cm²) is proposed to obtain such a low F1. We made some simulations to test this solution. In fact using an [a] oral tract configuration and opening the velum to the maximum we obtain F1=340Hz, F2=1000 Hz and F3=2055 Hz. These values are stable with pharynx variation.

Using a nasal model with a radiation area of 2 cm², like the one used in [7], F1 is as high as 428 Hz.

2.3 Interactive source model
The interactive source model is based in [1]. Several improvements were made: inclusion of a two-mass parametric model of glottal area; inclusion of Jitter, Shimmer and Aspiration; and capacity of synthesis of dynamic configurations.

2.4 Synthesis of varying tract shapes
As explained in the introduction, variation of the nasal coupling area with time seems to be an important characteristic of Portuguese nasal vowels. This makes necessary that the synthesizer may produce sounds resulting from time variable articulators. As the articulators vary in time, the impulse response of the system will also be altered. Ideally, we should obtain the tract shape and impulse response for each sampling instant (in this case sampling frequency is 10 kHz). This is computationally too intensive. Knowing that articulators evolve...
slowly, we consider the tract shape frozen during a pitch period and calculate the impulse response only at the beginning of each period. In our implementation, targets for each articulator can be defined independently. Articulator values between defined targets are obtained by linear interpolation. To improve the transition from two impulse responses, at beginning of two consecutive pitch periods, interpolation is performed with these two responses to obtain the impulse response at each sample instant.

2.5 Obtaining articulator positions
We need the tract shape for the vowels. At least for their beginning, where they are reported as oral. Direct measure of the tract shape is difficult and for the time being we don’t have, access to devices such X-rays, MRI, EMMA or others. The alternative, with inherent limitations, is to use an inversion method to obtain articulator positions from speech. For that, we choose an optimization process using the Simulated Annealing technique [6]. The process involves minimization of the difference between model and natural vowel formants.

3. EXPERIMENT - DYNAMIC VS STATIC
We will investigate if it is enough, to produce a good quality Portuguese nasal vowel, to couple the nasal tract or the degree of coupling variation in time is needed improves quality. The null hypothesis is that static and dynamic velum will produce stimuli of similar quality. If this hypothesis hold, listeners will not prefer one stimuli type against the other. On the other hand, if difference is statistically significant we can reject this null hypothesis, at a predefined confidence level, and have “proved” that the variation in time improves perceived quality. Our experiment consisted in a perceptual test using stimuli produced using the described articulatory synthesizer.

3.1 Method
3.1.1. Stimuli.
Stimuli were produced for the 5 nasal vowels using the articulatory synthesizer. Oral tract configuration was obtained using the inversion process already described. Resulting configurations were checked against bibliography reported configurations. Velum aperture was obtained manually by adjusting the ratio of nasal coupling area to oral area, in velum region, to a value of 10.

Timing, for the variable cases, was not very precise because we have no detailed production data. We used information in the literature about velum transition times and analysis of natural vowels in CVC contexts. We used the same timing for all vowels. In the first 100 msec velum stays closed, making an opening transition in 60 msec to the maximum value. Velum remains at this maximum. Stimuli with a nasal consonant, a bilabial ([m]), at the end are produced by closing the lips. Lips closing movement starts at 200 msec ending 50 msec later. Stimulus duration was fixed at 300 msec for all vowels. In the production of the stimuli, the interactive source model was used with variable F0. F0 starts around 100 Hz climbs 20 Hz in the first 100 msec and then gradually goes to 100 Hz. Other values for source parameters are presented in Table 1. Jitter and shimmer were added to improve naturalness. Tests, with oral vowels, showed that variable F0 and jitter, in particular, improved quality. Stimuli were generated by the synthesizer using 10kHz as sample rate.

For each of the vowels (5) we generated stimuli with dynamic and static velum. In the dynamic case two stimuli where generated: one with and other without a final nasal consonant. We get 3 stimuli for each vowel.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lungs Pressure</td>
<td>Pl</td>
<td>10000 dyne/cm²</td>
</tr>
<tr>
<td>Open Quotient</td>
<td>OQ</td>
<td>0.6</td>
</tr>
<tr>
<td>Speed Quotient</td>
<td>SQ</td>
<td>2</td>
</tr>
<tr>
<td>Minimum Glottal Area</td>
<td>Ag0</td>
<td>0.0 cm²</td>
</tr>
<tr>
<td>Maximum Glottal Area</td>
<td>Agmax</td>
<td>0.3 cm²</td>
</tr>
<tr>
<td>Steepness constant</td>
<td>Sk=A2-A1</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 1 - Values for source parameters used in the study.

3.1.2. Procedure
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.

Prior to application of the AB test, to know if listeners were able to discriminate the stimuli, a 4IAX discrimination test was performed by two listeners. Listeners were able to distinguish stimuli better than 95 %.

To conduct the test a computer program was developed to make the test completely automatic. Using as input the stimuli wave files and two configuration files the program makes the stimulus pairs, repetitions, and randomization, presents stimuli to subject and records test results for further analysis.

Signals were presented to the subjects over Sennheiser Headmax HD 470 headphones in a room with low ambient noise. Subjects responded by pressing, using the mouse, a button in the computer display.

The question asked to subjects was “Which of the stimuli do you prefer as a Portuguese nasal vowel?”

In preparing the test, we noticed that listeners had in some cases difficulty in choosing the preferred stimulus. This was caused by the two stimuli being of quality but too alike, or both being of poor quality. To handle this situation we added two new possible answers. We this we have four possible answers: “FIRST”, “SECOND”, “BOTH”, and “NONE”.

Stimuli were presented 5 times in both AB and BA order. Inter Stimuli Interval was 600 msec.

Test was divided in two parts. In the first part we compared static versus velum dynamic stimuli. In the second part comparison was made between dynamic stimuli with and without a final bilabial nasal consonant.

3.1.3. Subjects
A total of 11, 9 male and 2 female, European Portuguese native speakers participated in the test. Ages varied from 23 to 53 with mean 31.7 and standard deviation equal to 10.0.
3.2 Results

3.2.1. Listeners consistency and reliability
Results for each possible pair of stimuli used in the test were checked for listener consistency. Results were retained if listener preferred one stimulus in more than 60 % of the presentations. In this way, results of pairs in which listeners are no sure about their preference or consider the two stimuli alike or of poor quality are discarded. Only clear choices of one stimulus against other will be analyzed.

Listeners reliability [13] was examined by determining the judge to judge correlation. Due to test conditions reliability was low. Some listeners clearly responded to the test in an opposite way to the main trend. Despite this, we decide to keep all data.

3.2.2. Preferences
For the first part of the test, results of a correlated-samples t-test were significant, t(10)=2.236, p<0.05; stimuli using dynamic velum (Mean=39.55) were preferred to the stimuli with static velum (Mean=15.73). Results were not uniform for all the 5 vowels. We don’t present statistics by vowel due to the reduced number of listeners used.

In the second part of the test, listeners showed great difficulty in choosing. Most of the tests were inconsistent. We decide not to make statistical tests with these results.

4. DISCUSSION

Our null hypothesis, that stimuli with static and dynamic velum were perceived as having similar quality was rejected, with a 5 % significance level. Clearly, stimuli with dynamic velum are perceived as of higher naturalness. This points to the need of dynamic information to obtain nasal vowels of high quality. It is our main conviction that this influence of dynamics in perceived quality of European Portuguese nasal vowels deserves further study.

Considering our result, regarding the importance of dynamics to the perception of nasal vowels, new questions arise. One question is: what timing characteristics are preferred?

To address this question the idea is to produce stimuli with varying timings and have listeners choose their favorites. This is easier to say than done, because we easily arrive at a huge number of stimuli. To produce the stimuli and have them “ranked by preference” is a time consuming task. Despite this, with the help of an automatic program to produce the stimuli definition and a tournament strategy [Appendix A, 11] to perform the AB test the task could be reduced to a realizable task. We are starting these tests.

Further needed research is the direct study of velum position over time. For that techniques like EMMA and MRI could be of use.

ACKNOWLEDGMENTS
This work was made possible by the Ph.D. scholarship BD/3495/94 from the PRAXIS XXI. We also have to thank University of Florida for the stay of the first author at the MMIRC, headed by Professor D. G. Childers, where this work really started. We also thank Amália Andrade for making available more details of the work reported in [17].

NOTES
1. Portuguese language is very rich in nasal sounds. In European Portuguese, there are five nasal vowels, several diphthongs, triphthongs and three nasal consonants.
2. The main use of the interactive source will be in studies source-tract interaction effects in nasal vowels [22].

REFERENCES