RECOVERING GESTURES FROM SPEECH SIGNALS: A PRELIMINARY STUDY FOR NASAL VOWELS

Solange Rossato, Gang Feng, Rafaël Laboissière

Institut de la Communication Parlée Université Stendhal / INPG
Domaine Universitaire 38040 Grenoble FRANCE
email : rossato@icp.inpg.fr

ABSTRACT

For nasal vowels, a gesture as simple as the lowering of the velum produces complex acoustic spectra. However, we still find a relative simplicity in the perceptual space; nasality is perceived easily. In this preliminary study, we use statistic method to recover the gesture of the velum. In order to reduce the extreme variability of nasal vowels, we introduced a simulation based on Maeda's model instead of using a natural speech signal. In previous studies, nasality is supposed to increase either with size of the nasal area or with the area ratio between nasal and oral tracts at the extremity of the velum. In this work, both types of data are considered and analyzed with linear and non-linear tools. Finally, statistic inference is described and results are given for various areas of the nasal tract entrance and for various area ratios. The results show that velar port area is correctly estimated for small values while area ratio is a better parameter when velar port area increases.

1. ABOUT NASAL VOWELS

About 20 % of the UPSID languages have a phonemic contrast between oral vowels and nasal vowels. In French nasality is the only distinctive feature between the words [pɛ] *paix* (peace) and [pɛ~] *pain* (bread). Nasal vowels often derive from nasal consonant assimilation. This phenomenon also exists in languages with nasal consonants, more or less pronounced according to the context. Such vowels are called nasalized vowels in order to discriminate from nasal vowels. Nasalization, controlled or not, is a widespread feature.

On the articulatory level, vowel nasalization is produced by the lowering of the velum. This simple gesture connects the nasal fossa to the oral tract. The nasal fossa are quite complex and very different from one to another (Dang & Honda, 1994), but they are fixed for a given person. Acoustic consequences of this gesture depend on each person and also on the position of the velum. Many studies tried to find the acoustic correlates of nasality. Most of these agrees with a relative weakness of the first formant and proposes other secondary correlates. However detailed characteristics of the spectra of nasal vowels vary with the frequency of the first oral resonance and the magnitude of nasal coupling. In spite of such acoustic complexity, listeners give similar vowels nasality judgements, regardless of the phonological status of nasalization in their native language (Beddor & Strange, 1982).

Articulatory-to-acoustic relationship is quite complex and still not well understood for nasal vowels. Until now, there has been no corpus of velar port size measurements with simultaneous speech signal. The velum is an internal organ difficult of access. Nowadays IRM can very useful static vocal tract data. Furthermore, articulatory-acoustic inversion projects such as SPEECHMAPS have provided tools, data and encouraging results on oral vowels and fricative consonants. Can inversion techniques estimate the position of the velum or its evolution during nasal vowel production ?

In this preliminary study, we tried to answer this question in simplified and controlled conditions. Instead of natural speech signal, vocal tract simulation produces acoustic transfer functions. First, the production model and corresponding database are presented. Then database analyses are given before explaining in the third section the inversion technique applied and the results obtained.

2. PRODUCTION MODEL OF NASAL VOWELS

2.1. Articulatory Model

Usually articulatory models proposed by Maeda (1988) and Mermelstein (1973) consider only the oral tract. The model used in this work is Maeda's model based on X-ray images of a speaker (Patricia Barbier) pronouncing French sentences. Eight parameters regulate the jaw and tongue positions, the opening and the protrusion of the lips, the larynx length. A ninth parameter v_m is introduced to represent the velum position. Then, the articulatory model calculates the vocal tract area function whatever the value of the parameter v_m may be, as if the velar port was closed.

In Maeda's model, all parameters are normalised, centred on zero which corresponds to the mean value of all the positions registered on X-ray images and with a standard deviation 1. For normal distributions, variations between –3 and 3 are supposed to cover most of the cases. A more detailed study of the parameter v_m is needed to check that it has a normal distribution. The distribution of all the values of parameter v_m measured on X-ray images brings to the fore that this parameter does not follow such a distribution. The lowest values are around –1 and they are obtained during plosive consonant production. When parameter v_m is between –1 and 0, the velar port is closed. The highest values are close to 3 or 4 and correspond to the rest position, nasal vowels or consonants. The default value 0 is used for the oral vowel production, the velar port area is null. When parameter v_m increases, the velum lowers and allows the air to go through the nasal fossa. The velum is 1 cm thick and is ending by the uvula. Both outlines of X-ray image and midsagittal section obtained by the model are shown in Figure 1. So this articulatory model enables to get both oral area function and velar port area, areas being calculated from midsaggital sections. As data about Patricia Barbier's nasal

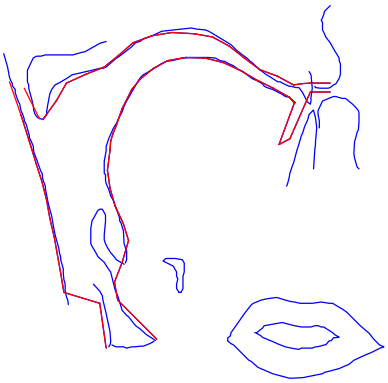

Figure 1 : Example of the outlines of a X-ray image and the outlines of the midsaggital section associated.

fossa are not available, area functions of nasal tract and sinuses are given by Feng and Castelli (1996). A set of nine parameters controls configurations of the vocal tract and corresponding area function. As this study only concerns vowels, constraints are imposed on the constriction area, the place of constriction and the lips area.

2.2 Construction of the Database

Traditionally simulation studies of nasal vowels compare transfer functions obtained with different velar port area, called hereafter nasal area A_n. The nasal area variations when parameter v_m increases are studied for various configurations. A bijective relation exists between the parameter v_m and the nasal area A_n except for high values of v_m where there is a saturation which depends on the tongue position. The velum lowers until it rests on the tongue and stays in that position even if parameter v_m still increases. Without taking into account such a saturation, the seven values corresponding to the nasal area values of 0, 0.2, 0.4, 0.8, 1.8, 2.2, 2.4, 2.6 cm² are determined. For a null value of parameter v_m, 1000 configurations are at random and transfer functions are computed for each of the seven values of parameter v_m. This database contains thus 7000 transfer functions.

However we cannot ignore nasal area saturation; for the same high value of parameter v_m, the nasal areas A_n can be very different because of the tongue position. To avoid such vowel dependence, Feng proposed to use another parameter: the area ratio d (Feng & Castelli, 1996).

$$d = \frac{A_n}{A_n + A_{oral}}$$

When the area ratio d is 0, the velar port is closed. When the area ratio is 1, vocal tract is a single-tube tract from glottis to nostrils, the pharyngo-nasal tract. When the area ratio is between 0 and 1, the nasal tract is connected to the oral tract with a certain degree of coupling. Pharyngo-nasal can be interpreted as a target towards which tend nasal vowels. The articulatory model enables us to obtain the oral configurations previously used in the first database. The velar port area is the area of the oral configuration at the extremity of the velum multiplied by the area ratio d. The oral tract area is multiplied by (1-d). Thus, the sum of both areas is constant. Transfer functions were computed for the area ratio values of 0, 0.05, 0.25, 0.5, 0.75, 0.95, 1. This second database depends on the area ratio and contains 8000 transfer functions.

3. DATABASE DESCRIPTIVE ANALYSIS

In this section, we attempt to establish an “axis of nasality”, i.e. where the nasal area A_n or the area ratio d increases. From a phonetic point of view, nasality can be considered as an additional dimension, perpendicular to oral space. Wright (1986) studied oral and nasal pairs in perceptual space and proposed the truncated cone hypothesis: loss of contrast between nasal vowels suggests a “conelike contraction” of the vocalic space boundary along the nasal dimension. Transfer functions are represented by 20 cepstral coefficients. Data analysis techniques provide a way to reveal the structure of those vectors or points of a 20-dimensional space.

3.1. Principal Component Analysis and Linear Regression

Principal Component Analysis (PCA) is a widespread and well-known data analysis technique to describe sets of points and reduce the dimensions of data. A 4-dimensional space preserves 90 % of oral transfer function variance and defines the oral subspace. Linear regression between nasal area values (or area ratio values) and cepstral coefficients provides a “nasality axis”. This axis is nearly perpendicular to the oral subspace (86.77° instead of 90°). For each database, points were projected on this “nasality axis” and the repartition is shown in Figure 2. The nasal areas A_n of 0, 0.4 and 0.8 cm² are well distinguished. For higher values, curves are superposed. On the contrary, projections have different mean values for an area ratio d of 0.5, 0.75 and 1.

3.2. Curvilinear Component Analysis

Curvilinear Component Analysis (CCA) is a self-organised neural network that performs an efficient dimensionality reduction preserving the underlying structure of the data (Desmartines & Herault, 1997). An interesting property of this new technique lies in the possibility of introducing a priori information to constrain the input-output mapping on its initialization. The output space dimension for oral cepstral coefficient vectors is 4, as for PCA. The fifth dimension, perpendicular to the oral space, is initialized according to the nasal area values (or the area ratio values). After learning, CCA connects each point of the 20-dimensional input space to a point in the 5-dimensional output space.

First, vector quantization reduces the number of vectors in the database in order to keep the general structure of the database and bound computing time. Then, the self-organised neural network learns the reduced database and locates each input point in the output space preserving to the maximum small distances. Next, another neural network realizes a continuous projection that associates to each input vector its image in the output space.

Projections along the fifth dimension give curves very similar to those obtained with linear regression. These results confirm that a variation in velar port size leads to important changes in transfer function for small areas only, thereafter area ratio seems

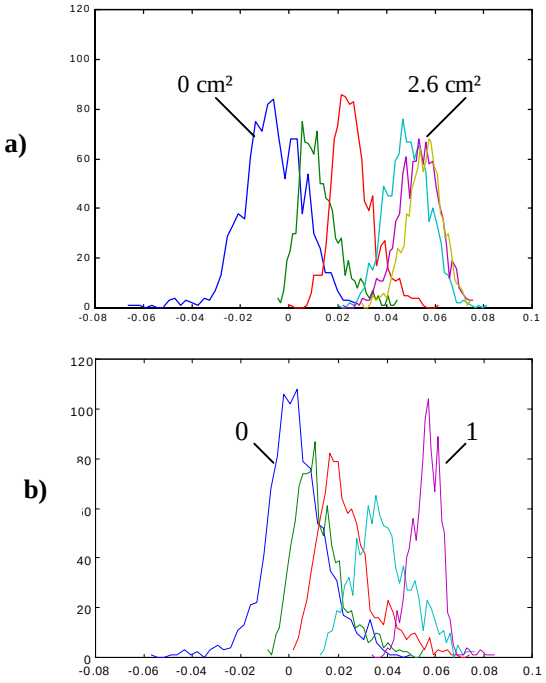

Figure 2: Projection on the "nasality axis" provided by linear regression for **a)** nasal areas 0, 0.4, 0.8, 1.8 2.2 and 2.6 cm² **b)** area ratios 0, 0.25, 0.5, 0.75 and 1.

to be more relevant in distinguishing two levels. It would be interesting to compare both the CCA fifth axis and the linear regression axis but CCA gives no relation between input and output space.

3.3. Large-Scale Integration

Listeners have several perceptual representations of a given sound. One of them involves a very large-scale integration (3 Barks) of the acoustic spectra (Chistovitch & Lubinskaya, 1979). Both lowest peaks, called F'1 and F'2, are extracted from the integrated transfer function. For oral vowels, F'1 and F'2 are characteristic features of the vowel. This integration was applied to nasal vowel transfer functions. For each database, extremes were plotted on the F'1-F'2 plane as shown in Figure 3: peaks of the oral transfer functions and peaks of transfer functions with a nasal area A_n of 2.6 cm² (or less if saturated); oral peaks and peaks of pharyngo-nasal transfer functions (area ratio equals 1).

The first remark consists in pointing out that transfer functions with a nasal area A_n of 2.6 cm² and pharyngo-nasal transfer functions are all located in the right top corner of the vocalic "triangle". In Figure 3b), the pharyngo-nasal target is clearly present and close to the 300-1000 Hz proposed as nasal peaks by Maeda (1984). Most of pharyngo-nasal configurations are situated near 300 Hz for F'1 and 800 Hz for F'2. Some of them are very far from this area. Indeed with automatic detection 2 very close peaks are detected as a single one and thus the peaks plotted are F'1 and F'3 or F'2 and F'3. These points are peaks of pharyngo-nasal tracts of oral configuration close to [u]. For a nasal area A_n of 2.6 cm², peaks are closely grouped around 350-900 Hz as shown in Figure 3a). Regarding their articulatory configurations, we realize that they are nearly pharyngo-nasal tracts: the velum is so low that it closes the oral tract. Some points are scattered and correspond to front vowels with high or middle aperture, when the velum can go low enough to have 2.6 cm² velar port area without closing the oral tract.

If, in both cases, masses are associated to pharyngo-nasal tracts, why do they not represent the same mass ? Both databases were created in different ways. For the first one, the nasal area A_n and oral area function are calculated with the model including the velum. For the other database, area functions are inferred from oral vocal tract and area ratio by linear relations. Values of velar port area cannot correspond whether we introduce the simulation of the velum or not.

4. INVERSION AND FIRST RESULTS

The aim of inversion is to estimate from a transfer function H a suitable parameter (called *v*) indicating the position of the velum. The probabilistic approach is a global method which does not need specific knowledge about the direct system. It lies in learning, its performance depends on the database which must be the most representative to be able to give correct probabilities for further samples.

4.1. Probabilistic Inference

The estimated parameter $\hat{v}$ will maximize the probability *p(v/H)*.

$$\hat{v} = \arg\left(\max_{v} \; p(v / H) \right)$$

The problem is to determine the probability *p(v/H)*. Bayesian rules can reverse terms and we have:

$$p(v / H) = \frac{p(H / v) \cdot p(v)}{p(H)}$$

where *p(v)* is the *a priori* probability of having parameter *v* and *p(H)* is the *a priori* probability of having transfer function H. We have no peculiar knowledge of these probabilities and there is no justification for selecting one value of parameter over any other; transfer functions are also taken to be equiprobable. So the probabilities *p(v)* and *p(H)* are uniform, *p(v/H)* and *p(H/v)* are proportional. The estimated parameter maximizes the likelihood *p(H/v)*.

The parameter *v* is sampled and the probabilities *p(H/v)* for each value of *v* are normal distributions whose mean and covariance matrix are estimated from the learning database.

4.2. Implementing Probabilistic Inference and Results

Transfer functions are represented by a set of 20 coefficients which are obtained either from Fourier transform (cepstral –CC– and mels cepstral –MFCC– coefficients) or from linear prediction (LAR and LSP coefficients, used in coding). The parameter indicating the velum position is either the nasal area A_n or the area ratio *d* according to the learning database. 70 % of each database constitutes the learning database, the 30 % remaining is reserved for testing. The estimated value of the nasal area A_n (or the area ratio, respectively) is compared with

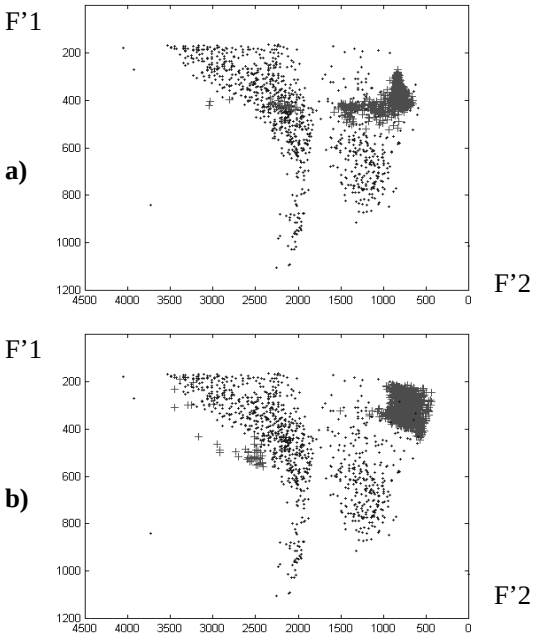

Figure 3: F'1-F'2 plane for **a)** nasal areas 0 (points) and 2.6 cm² (cruces) **b)** area ratios 0 (points) and 1 (cruces).

the effective nasal area A_n (or the area ratio, respectively) and the correct estimation rate is shown in Figure 4. The estimation rate is excellent for small values of the nasal area, until 0.8 cm² whatever the set of coefficients. It falls for nasal area higher than 1.8 cm². This limit can be explained by the tongue saturation that begins for nasal areas larger than 1 cm². For the second database, correct estimation rate is good whatever the value of area ratio, excellent for high values. For small area ratio (small opening of the velar port with regard to the oral tract entrance), coefficients obtained by linear prediction give better estimation rates; linear prediction seems to be more appropriate for representing vocal tract resonance than classical Fourier transform. When the area ratio increases, all coefficients give similar estimation rates.

CONCLUSION

With this simulation model, we can conclude that the estimation of small nasal areas is very good and that as the velum is lowered, the area ratio becomes a better parameter than nasal area. Acoustic spectra seem to be sensitive to velar port area constriction ; when the area increases, constriction disappears and the area ratio between the oral and nasal tracts is more representative of the changes in transfer functions. Further studies on natural speech signals are needed to validate this probabilistic method that provide good results in simulation cases.

ACKNOWLEGMENTS

The authors wish to thank Jean-Luc Schwartz for our stimulating discussions throughout this work and for reviewing earlier versions of this paper. We are also grateful to Louis-Jean Boë and Nathalie Vallée for advice on the articulatory model.

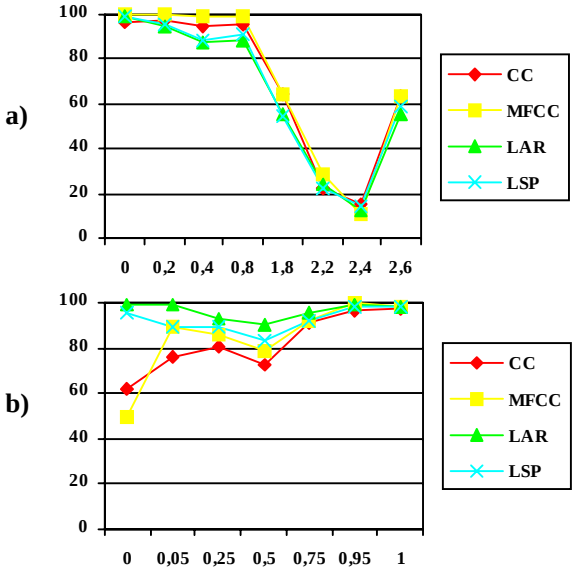

Figure 4: Correct estimation rate **a)** nasal area database **b)** area ratio database.

REFERENCES

1. Beddor P.S. and StrangeW. "Cross-language study of perception of the oral-nasal distinction," *J. Acoust. Soc. Amer*. 71, 1551-1561, 1982

2. Chistovich L.A. and Lublinskaya V.V. "The center of gravity effect in vowel spectra and critical distance between formants," *Hear. Res.* 1, 185-195, 1979.

3. Dang J., Honda K. and Suzuki H. "Morphological and acoustical analysis of the nasal and the paranasal cavities," *J. Acoust.. Soc. Amer*. 96, 2088-2100, 1994.

4. Desmartines P. and Hérault J. Curvilinear Component Analysis: a self-organizing neural network for nonlinear mapping of data sets, *IEEE Trans. Neural Networks* 8: 148-154, 1997.

5. Feng G. and Castelli E. "Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization," *J. Acoust. Soc. Amer*. 99, 3694-3706, 1996.

6. Maeda S. "Une paire de pics spectraux comme corrélat acoustique de la nasalisation des voyelles," *13éme J.E.P.*, Bruxelles,223-224,1984.

7. Maeda S. Improved Articulatory model. *J. Acoust. Soc. Amer*. 84, S146, 1988.

8. Mermelstein P. "Articulatory model for the study of speech production," *J. Acoust. Soc. Amer*. 53, 1070-108, 1973.

9 Wright J.T "The behavior of nasalized vowels in the perceptual vowel space," *Experimental Phonology*, edited by John J. Ohala et Jeri J. Jaeger (Academic Presse, INC), 1986.