EXTRACTING NASALITY FROM SPEECH SIGNALS

Henning Reetz

Max-Planck-Institut für Psycholinguistik, The Netherlands

ABSTRACT

Nasality in this paper refers to voiced nasal consonants, nasal vowels, and nasalized vowels produced by non-pathologic speakers. An algorithm is presented, that segments the speech waveform into *nasal* and *non-nasal* parts. The decision whether nasality exists is based on two features of the speech signal: (1) Presence of vocal cord vibration (voicing) and (2) presence of a resonance in the range 200 Hz - 400 Hz. The frequency of this resonance may vary between speakers but is constant for a single speakers. The algorithm is validated by comparing it to results of a perception experiment. The problems to define and measure nasality are sketched before the algorithm is presented.

1. INTRODUCTION

Nasality is defined in different ways in different fields. It can be determined by underlying or surface structures of the speech material, it can be described by articulatory movements or physiological data, it can be related to signal characteristics that are expected on the basis of theoretical considerations, and it can be determined in perception tests where subjects use their native or experience listening skills. Each field defines *nasality* within its system and the definitions must not necessarily coincide. For example, [10] reported velar opening on (phonetically) non-nasal sounds and [6] found coupling to the nasal cavities even if the sound was not perceived as nasalized by listeners.

In addition to the problem how to define nasality comes the problem how to measure it. In articulatory experiments it is a complicated task to determine the position of the velum that lies hidden in the back of the mouth. Physiological experiments have to measure (nasal) air flow accurately without effecting normal articulatory behavior. Acoustical analyses have to separate properties of the nasal system from the oral and pharyngeal system. In perception tests it is not trivial to create listening conditions that are similar to everyday speech perception situations. If in any field indirect measurements are performed, it can happen that the coherence between the measured quantity and the related quantity does not exist in the expected way. For example, nasal airflow is not necessarily a measure for the size of the velar opening and it has been reported that changes in velar heights occur even though the velar port is closed ([2]).

In this paper an acoustical measure will be used that intends to be related to perceptual nasality. First we will brief some acoustical properties of nasal sounds and ways to measure them. In the following chapter an algorithm to extract low-frequency resonance will be presented. After describing a perception experiment the relation between the low-frequency resonance and the perception results is demonstrated and discussed.

2. SOME ACOUSTIC PROPERTIES AND MEASUREMENTS

Acoustic theory predicts extra anti-resonances and resonances for the additional shunting branch of the nasal cavity system ([3]). These effects can be found in the speech signal, though they are not easy to identify ([3]). Common acoustical features found in the literature are low-frequency energy and broadening of the overall spectral pattern ([4]). The occurrence of a low-frequency resonance can be explained by a Helmholtz resonating cavity. [7] observed a resonating frequency of the sinus maxillares in the range of 200-800 Hz. [1] computed a resonating cavity of about 800 cm³ for a 250 Hz resonance and proposed (parts of) the scalp as possible resonator. The flattening of the spectrum can be explained by the complicated nasal cavity structure inducing many partial resonating and shunting cavities ([7]).

Spectrographic (or FFT) analyses have the advantage to preserve all information of the acoustical signal. This can be a handicap because the ground frequency (and the window size) is apparent in the spectrogram. Smoothing these spectra reduces the appearance of these effects but also reduces the spectral resolution. Source-filter models separate the properties of the source signal (F0) from the oral and nasal tract, but add assumptions about the model to the estimated spectra. All-pole LPC does not include shunting cavities in its model and therefore seems to be inadequate. Fitting the model to the analyses purposes (i.e. determining the number of poles) is done by rules of thumb. Spectral estimation methods that include zeros in their models (like analyses-by-synthesis schemes or ARMA models) inherit the problem to estimate the number of poles and zeros. A problem that is non-deterministic because a zero can be approximated by a large number of poles ([8]).

In this study the low-frequency resonance will be used as a gross indicator for a phenomenon that listeners would call *nasality*. This feature is reported in many articles studying the acoustical properties of nasality and seems to us more obvious in the spectrum than the theoretically more founded existence of zeros. We are not interested in detecting zeros at all and choose LPC analyses as the spectral estimation method. This has several advantages: it is easy to compute, it is commonly available, and it suppresses the effect of the ground frequency in the estimated spectrum.

3. ALGORITHM

In a pre-processing step the speech signal is separated in voiced and unvoiced parts using a pitch extraction algorithm ([11]). This reduces the amount of computing to be performed in the following steps and excludes background or friction noise that might show low-frequency prominence.

In the next step high order autocorrelation LPC analyses with parabolic interpolation is carried out on the voiced parts (25.6 ms Hamming window, 12.8 ms step rate). The order of the LPC must be adjusted manually to the recording and voice quality. For a 10 kHz recording, values between 14 and 28 poles were found to practical.

In the last step a low-frequency resonance is searched that lies below 400 Hz, is constant in frequency, and lasts for at least 40 msec. Constant in frequency means here that the center frequency of the resonance does not vary more than 10% in the adjacent frames.

4. EVALUATION

The performance of the algorithm is compared to the results of a lexical decision experiment carried out by [5]. The study included a gating experiment on British English monosyllabic words of the form CVC and CVN. Listeners were presented gradually incrementing information (gates) from word-onset until the entire word was heard. The listeners' task was to write down the word they were hearing.The length of the gates were incremented by about 40 ms from step to step. One gate of each word were set at the offset of the vowel. 20 minimal CVN-CVC pairs (40 words) spoken by one male speaker

were presented with 9 to 13 gates giving a total of 441 gates. Details of the experimental procedure can be found in [5].

The responses of the subjects were classified in the following manner (this classification was not part of the experiment of [5]): If more than 50% of the listeners wrote down a CVN word for a gate then it was called nasal, otherwise oral. Because the listeners heard gates with increasing length, it happens that a gate was judged oral, while the next, longer gate, was judged nasal (or vice versa). By this it is possible to make a statement about the 40 ms incremental part of a gate whether it were or were not perceptually nasal. We will name an incremental part of a gate and the first gate of a series *segment* from now on.

The speech recordings used in the perception test were analyzed by the described algorithm as well. If the bandwidth of the low-frequency resonance as found by the algorithm was below 150 Hz the sounds were classified as nasal, otherwise as non-nasal. If the results of the perception test and the algorithm for a segment did not coincide it was counted as an error.

Of the 441 segments, 322 were perceived as nasal and 119 as oral. 67 (15%) of the segments were mis-labelled by the algorithm. 50 segments perceived as oral were classified nasal, in 17 cases a perceptual nasal segment was classified oral by the algorithm. We will first characterize the errors of the 50 *false alarms* then the 17 *missings*.

Of the 50 *false alarms*, 10 were segments where the algorithm classified them *too early* as nasal in front of perceptually nasal segments. 9 segments were clustered in 2 cases were a vowel in front of a nasal consonant was perceived as oral by the listeners. 29 times classified the algorithm the pre-voicing of voiced stops as nasal. And 2 segments of /l/ in CLAD were classified by the algorithm as nasal.

Of the 17 *missings* were 4 segments, where the algorithm classified them *too late* as nasal at the beginning of perceptually nasal segments. The other 13 nasal segments were missed in 4 clusters inside oral or nasalized contexts: in /l/ in CLOWN, and in the vowels of BRAN, VAIN, and TRADE.

5. DISCUSSION

The errors made by the algorithm can be grouped in 4 categories: (1) *too early* or *too late* indication of nasality, (2) judgment of pre-voicing as nasal, (3) mis-classification of /l/, (4) and mis-classification of nasalized vowels.

The first category of errors can be taken into account as jitter between perceptual and acoustic segmentation.

The second category - the classification of pre-voicing as nasals - is an interesting result. Airflow measurements have observed nasal airflow at the beginning of pre-voicing, either due to insufficient velar closure together with oral closure and increasing intra-oral pressure, or due to an upward push of the uvula extinguishing air from the nasal tract. Whether a nasal coupling exists and in how far a perception test without context information will yield nasal responses in these cases stays unclear.

The third category - mis-classification of /l/ - appears in both directions. This is a frequently mentioned effect ([9]).

The mis-classification of nasalized vowels reflects the more complicated structure of these sounds. It should be taken into account that 50% of all words contained (phonetical) nasalized sounds and that in most of these cases the results of the algorithm went along with the results of the perception test: When the listeners heard a word as nasal from an early stage of the vowel, the algorithm did so. When the listeners perceived the nasality at the end of such a vowel, so did the algorithm. Only in 4 cases we found a mis-labelling.

6. CONCLUSION

The algorithm uses a very gross feature to identify nasality ignoring context effects and formant transitions as used by [9]. He found the mean centroid frequency below 500 Hz to be the most useful contributor to the total categorization score. He observed further that a value for the first resonance near 250 Hz is a necessary but not sufficient condition for the existence of nasal murmurs (because this property is shared by the first formant frequency of high vowels). Most of his false indications result from the confusion of liquids, glides and semivowels with nasals.

The simpler algorithm used here shares obviously some of these results. But to come to a stronger statement about the behavior of the proposed method and to get more evidence for the applicability of the algorithm as a valid nasality classificator more investigation will be performed. A larger set of words of different speakers including a significant set of nasal vowels must be tested. An open question is, whether the algorithm measures a low-frequency resonance originated by the coupling of nasal tract, or whether F0 or a harmonic of it is detected. [4] and [1] observed this resonance independent of the F0 frequency and informal tests with the proposed algorithm yielded the same result. Analyses of high pitched voices have given mixed results: some voices showed good results, while other voices were classified as nasal for most of the voiced segments. Whether these are only artifacts of the algorithm or whether those voices are nasal has not been proved by perceptual tests. Application to running speech gave the impression that nasal consonants and voiced stops are consistently detected by the algorithm.

We are aware of the fact that the algorithm does not have any normalizing mechanisms and does not perform any analyses of transitions. To some extend these features could be modelled by an adaption strategy for the number of poles and an investigation of the development of the bandwiths in time. But we first want to study the algorithm at the present state and collect experience about its behavior before extending it.

7. REFERENCES

[1] Castelli, E., Perrier, P., Badin, P. (1989), "Acoustic considerations upon the low nasal formant based on nasopharyngeal tract transfer function measurements", *European Conference on Speech Technology, II*, Paris, 412-415.

[2] Counihan, D. T. (1979), "Oral and nasal airflow and air pressure measures", In: Bzoch, K.R. (Ed.) *"Communicative disorders related to cleft lip and palate"*, Little, Brown & Company, Boston, 269-276.

[3] Fant, G. (1960), *"Acoustic theory of speech production"*, Mouton & Company, The Hague, Netherlands.

[4] Fujimura, O., Lindqvist, J. (1971), "Sweep-tone measurements of vocal-tract characteristics", *JASA, 49*, 541-558.

[5]Lahiri, A., Marslen-Wilson, W. (1991), "The mental representation of lexical form: a phonological approach to the recognition lexicon", *Cognition (in print)*.

[6] Lindqvist, J. (1965) "Studies of the voice source by means of inverse filtering" *Speech Transmission Laboratory - Quaterly Progress and Status Report, 2*, Stockholm, 8-13

[7] Lindqvist-Gauffin, J., Sundberg, J. (1976), "Acoustic properties of the nasal tract", *Phonetica, 33*, 161-168.

[8] Makhoul, J. (1975), "Linear Prediction: a tutorial review", *Proceedings of the IEEE, 63*, 561-580.

[9] Mermelstein, P. (1977), "On detecting nasals in continuous speech", *JASA, 61*, 581-587

[10]Moll, K.L. (1962), "Velopharyngeal Closure on Vowels", *Journal of Speech and Hearing Research, 5*, 30-37

[11]Reetz, H. (1989), "A Fast expert program for pitch extraction", *European Conference on Speech Technology, I*, Paris, 476-479.