On the Detection and Classification of Frames from European Portuguese Oral and Nasal Vowels

Yosvany Llerena Rodriguez[1], *António Teixeira*[2]

[1]IEETA, Universidade de Aveiro, Portugal
[2]DETI/IEETA, Universidade de Aveiro, Portugal
llerena@ua.pt, ajst@ua.pt

Abstract

Our aim is to perform a comparative evaluation of potential of oral versus nasal sounds in a European Portuguese Speaker Verification system. For that, we report, in this paper, the work on the necessary detection of the relevant segments. Implemented detection and classification consists in a typical cascade of speech framing, feature extraction and the use of classifiers. A total of 31 different features - including a subset of the features used recently by Pruthi and coworkers for American English nasal vowels detection - were extracted from each frame. Taking into account the small dataset restriction, we selected three classifiers: the well known SVM; the, more recent, Naive Credal Classifier 2 (from the Naive Bayes family of classifiers) and a Metaclassifier based on boosting (MultiBoostAB). Results, using a small database, showed as the best classifier the MultiBoostAB. Best results for Recall, Precision and F-measure, of 87.04, 88.0 and 87.5 %, were obtained for this classifier when trained with an equal number of samples of each class and non-including the first 40% of the production of the nasal vowels.

Index Terms: Nasal vowels, Blind Segmentation, Naive Credal Classifier 2, MultiBoost, European Portuguese.

1. Introduction

Segmentation and labeling of speech material according to phonetic or similar linguistic rules is a fundamental task in speech processing. A "Blind speech segmentation procedure allows a speech sample to be segmented into sub-word units without the knowledge of any linguistic information (such as, orthographic or phonetic transcription) [1].

One application of speech segmentation is on Speaker Verification systems, to provide a sub-word level segmentation [1].

Nasal sounds are more speaker-dependent due to the considerable differences among individuals shapes of the resonators involved [2]. These differences on size and shape of the nasal cavities can cause differences in the spectral characteristics of nasal murmurs for different speakers [3]. As Portuguese is a language rich in nasal sounds, the segmentation, detection and classification of these sounds needs attention.

According to [4] a vowel nasalization detector is essential for speech recognition (particularly the recent landmark-based recognition) in languages with phonemic nasalization (as is the case of Portuguese). To best of our knowledge no such detector has been developed yet.

Our goal is to perform a comparative evaluation of oral versus nasal sounds potential in an EP speaker verification system. For that, we report, in this paper, the work on the necessary detection of the relevant segments. As no comparable database as the TIMIT used by, for example, Pruthi [5], for our investigations, the problem of developing the speaker independent segmentation system included the constraint of being developed with only a small amount of data (pre-existent if possible), including noisy recording conditions e less than perfect articulations.

1.1. Related work

Recently the detection/segmentation of nasal vowels was addressed by Pruthi and coworkers. In [5], the authors evaluate, for American English, a set of nine acoustic parameters (APs). Those APs came from previous studies and almost referenced as capable to well describe the vowel and nasalize segments. In addition, they work only with the middle 1/3rd of the frames for oral vowels and the last 1/3rd of the frames for nasalized ones. They used SVMs with linear and Radial Basis Function (RBF) kernels as classification methods and three different databases (StoryDB, TIMIT and WS96/97). The SVM outputs were mapped to pseudo-posteriors histogram to achieve the final decision using a probabilistic measurement. Results for selected APs were compared with other two sets (6 other features and 39 MFCC). Best accuracies were achieved by the RBF Kernel SVM with proposed APs, obtaining accuracies of 96.28%, 77.90% and 69.58%, respectively for the databases StoryDB, TIMIT and W96/97.

Most representative works on Portuguese phoneme segmentation were performed using HMMs, not suitable for small datasets situations, such as the one addressed in this paper.

2. System Overview

The speech waveform is first split into small segments (frames). All subsequent processing is frame based.

First step consists in feature extraction (details in sec. 3) for each of the frames of the input signal. As our target classes are included in the Voiced part of the speech signal, our second step is to classify each frame as Voiced or Unvoiced. The algorithm used for this step is based in [6]. As a third step, classifiers are applied to classify each frame in one of the classes of interest: Oral Vowels, Nasal Vowels, Other. The classifiers considered were chosen by their potential to being trained with a small data set.

3. Features

An extended set of 31 features was used in our experiments. A subset of the features used recently by [5] was implemented, being the basis of our feature set. Our purpose is to investigate if they can be useful for languages other than English, particularly in European Portuguese (EP) - a language with nasal vowels in its phonological inventory. Due to their higher relevance for the present work, these features are described in the following subsections. Other feature groups were also included: frequencies (F0,F1,F2,F3,F4,Instant Frequency), Childer's [7] nasal and vowel detection features (Nasal Rate, Vowel Rate, Volume of Low and High Band frequencies) and others commonly used in speech segmentation (Energy, Energy rate, Energy in low and high bands, ZCR, Entropy, Spectral features (rollOff, centroid, flux, flattening)).

3.1. teF1

Teager energy operator was used by [8] for hypernasality detection, using a pitch synchronous approach. As an alternative [5] proposed the use of the correlation between the Teager energy profiles of lowpass filtered speech and bandpass filtered speech, the correlation between the Teager energy profiles of narrow bandpass filtered speech and wide bandpass filtered speech centered around two different frequency regions be considered. In this case, the frequency regions were centered on the first two formant frequencies, obtained from a formant tracker.

Teager energy profile, $\Psi_d[x(n)]$, for a signal $x(n)$ is calculated as:

$$\Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad (1)$$

The teF1 feature can be estimated as the correlation between Narrow and Wide band filtered signals:

$$teF1 = \rho(\Psi_d[sNBF1], \Psi_d[sWBF1]) \qquad (2)$$

where sNBF1/F2 and sWBF1/F2 are the narrowband and wideband filtered speech signal centered around F1/F2 respectively. The narrowband filter uses a 100 Hz bandwidth and the wideband filter was set to 1000 Hz both with a 200 order filter.

3.2. nPeaks40dB

This feature [5] was designed to capture the large extra poles across spectrum as author mention. Basically count the number of peaks in limits of 40 dB of maximum dB amplitude of complex cepstral spectrum. Including only peaks between $0 - 4000$ Hz.

3.3. F1BW

Pruthi in [5] suggest that even though bandwidths of oral formants may not increase due to the losses in the nasal cavity, the bandwidths of these formants may seems to be wider because of unresolved poles which appear at frequencies very close these oral formants. The feature in question means is obtained from the ESPS formant tracker algorithm, as made available in the Snack ToolKit.

3.4. A1-H1fmt

In [9] author distinguish the difference between the amplitude of the first formant (A1) and the first harmonic (H1), and change in A1-H1 over time as being correlated to the perception of nasality. A reduction in A1-H1 is expected because A1 reduces with nasalization also confirmed by [5].

3.5. std0-1k

Standard deviation (STD) of the local spread of energy around the centre of mass (CM) was found to be a very good measure of nasalization. The feature can be estimated by measuring the second moment of local energy around CM. The term "local" was defined to include all energy within a specified frequency radius of CM [10].

Calculated between two frequency ranges, f_1, and f_2, the CM, $\bar{f}$, is defined as:

$$\bar{f} = \frac{1}{A_1}\sum_{f=f_1}^{f_2} fX(f) \text{ with } A_1 = \sum_{f=f_1}^{f_2} X(f) \qquad (3)$$

where $X(f)$, is the value of the Discrete Fourier Transform (DFT) spectra at frequency f. For nasalized vowels the CM must be computed between 0 and 1 kHz, which covers the first formant range of most men and women.

Before STD estimation and proposed by Glass in [10], the DFT spectra was windowed with a trapezoidal windows before the CM computation to reduce the CM function sensitivity to sudden changes at the end points, such as formant passing below 1000 Hz. The windows shape is flat between 100 and 900 Hz, and had 100 Hz tapers at each end. Applying this windows the spectra were not sudden changes in CM caused by the marginal movement in energy across the upper boundary.

In [5] the author also proposes, before CM calculation, to set any amplitude value less than threshold (20 dB below maximum) equal to threshold, and then subtract threshold from all values to set floor to zero.

4. Classifiers

Taking in consideration the small dataset restriction, as classifiers, besides the commonly used SVM, 2 others were selected. One from the Naive Bayes family, the Naive Credal Classifier 2 (NCC2) [11]; the other a Meta classifier using Boosting [12].

4.1. Support Vector Machines (SVM)

SVMs (ex: [12]) are based on the concept of decision planes that separates between a set of objects having different class memberships. Most classification tasks demand complex structures in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). To address this problem, the original objects are mapped using a set of mathematical functions, known as kernels. The mapped objects will be linearly separable and, thus, instead of constructing the complex separation curve, all we have to do is to find an optimal line that can separate the classes. There are number of kernels that can be used in SVM models. These include linear, polynomial, RBF and sigmoid. The RBF is by far the most popular choice of kernel types. This is mainly because of their localized and finite responses across the entire range of the real x-axis. We used libSVM implementation, running under Weka [12].

4.2. Naive Credal Classifier 2 (NCC2)

The NCC2 is an extension of Naive Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with **small or incomplete data sets** [11]. Robustness is achieved by delivering set-valued classifications (that is, returning multiple classes) on the instances for which (i) the learning set is not informative enough to smooth the effect of choice of the prior density or (ii) the uncertainty arising from missing data prevents the reliable indication of a single class. As on small data sets Naive Bayes Classifiers (NBC) may return prior-dependent classifications, leading to fragile predictions, to deal with this problem, NCC2 specifies a set of prior densities, referred to as prior credal set; the credal set is then turned into a set of posteriors via element-wise application of Bayes rule [13].

We used the Java implementation, named JNCC2, released under the GNU GPL license and capable of processing ARFF format files.

4.3. AdaBoost

AdaBoost, a diminutive for Adaptive Boosting [14], is an algorithm for creating a "potent" binary classifier as linear combination of a simple one. Boosting decides the weak classifiers and their weights based on the minimizing of loss function in a two-class problem. Boosting is usually fast and has high performance. As an interesting property of Adaboost, we can mention the potential to reduce bias and variance from, for example, tree based classifiers.

For the present work we used the MultiBoostAB made available in Weka [12].

5. Classification experiment

This experiment evaluated the efficiency of the proposed system. Given a set of features extracted form EP speech frames, the system had to assign a class to each of the frames. The classes considered were: Nasal Vowel, Oral Vowel and Others.

5.1. Database

A small database was created consisting of 2 parts:

First part consisted on 3 minutes of speech from an EP native speaker reading random news. Recording took place in a normal office. Speech signal was recorded using 22050 Hz sample rate at 16 bit mono.

The second part consists on the speech recordings made during ElectroMagnetic Midsagittal Articulography (EMMA) acquisition for EP nasals. It is very rich in nasal sounds, particularly nasal vowels, and an example of noisy speech produced with articulations far from perfect. The database includes recordings from two native EP speakers(one male and one female). Two speaking rate conditions were recorded: normal and fast rate.

First part was manually annotated at segment level for all produced sounds; the second part was not fully annotated: only oral and nasal vowels segments and their context were contemplated.

Considering the frames (20 ms, no overlap), the database has the following distribution: 6836 Oral Vowels, 9763 Nasal Vowels and 43236 Others. If only the final 60% frames from nasal vowels are kept, the number of frames from nasal vowels decreases to 3377.

5.2. Metrics

For evaluation, we used 3 criteria: recall ratio (R), precision ratio (P) and F-ratio (F): $R = \frac{tp}{tp+fn}$, $P = \frac{tp}{tp+fp}$, $F = \frac{2RP}{R+P}$, being tp the number of true positives, fp the false positives, fn the false negatives.

As we are only interested on two of the three classes (Nasal and Oral vowels), values for R, P and F reported are averages of this metrics calculated separately for each of the two classes.

5.3. Results

5.3.1. Classifiers comparison

We started our experiments by comparing the three classifiers in a common setup, 10 fold cross validation. Besides classifier effect, two other factors were considered: the balanced number of examples for each class and the use or not of all the frames from the nasal vowels. This resulted in 4 evaluation scenarios. The inclusion of the second factor (discarding the initial frames of nasal nasal) was motivated by reports claiming that they have an initial oral (or oral like) phase. For NCC2 only determinate classification (only one class selected) is considered. The results are presented in Fig. 1.

From the figure is clear that the best results were obtained with the MultiBoost classifier. The SVM gave, in general, poor results. For all classifiers the best performance was obtained when combining the use for training of a balanced number of samples of each class and the inclusion of only the frames from the final 60% of the nasal vowels productions.

5.3.2. Evaluation of features subsets

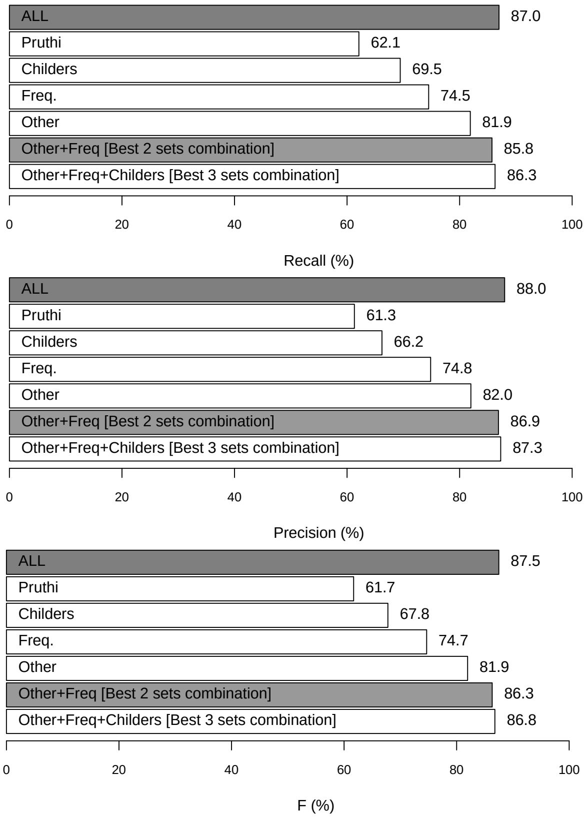

Figure 2: Results of the evaluation with the MultiBoostAB classifier on different subsets, and combinations of subsets, of features. In all cases the training was performed with equal number of examples for all classes and discarding initial part of nasal vowels production.

The results for evaluation on different subsets of the features (maintaining the results for ALL features as reference) are presented in Fig. 2. Results on each of the four subsets mention in 3 are complemented with the best results obtained on two and three subsets combinations.

The best results for a single subset were obtained with the "Others" subset, followed by the "Frequency" subset. Contrary

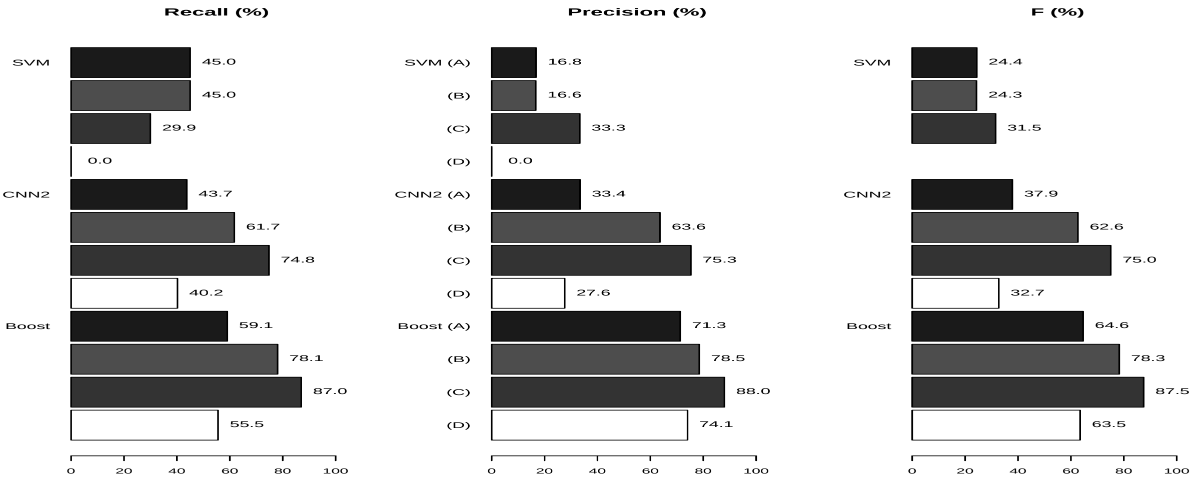

Figure 1: Results for Recall, Precision and F-measure for the 3 classifiers when evaluated with all the features. The following test variants were considered: all frames used (unbalanced number of samples for each class), named A; number of samples equal for the 3 classes (B); balanced with only the final 60 % frames from nasal vowels (C); unbalanced version of the previous (D).

to our expectations, the worst results were for the "Pruthi" subset. As best results with only one subset is more than 5 % lower than using all features, combinations were also evaluated. The two best subsets work well when combined and constitute the best combination of two subsets (Others+Frequency). The best 3 subsets combination is also formed by the classifiers on the first 3 places when used alone. With these subsets the performance is quite close to the one obtained with all features, interestingly not including features of the "Pruthi" subset.

6. Conclusion

In this paper we investigated the usefulness of three classifiers and several subsets features in the detection and classification of speech frames from productions of EP oral and nasal vowels. The classic SVM classifier was compared against the NCC2 and MultiBoost classifiers. A small data set was used for classifiers and features evaluation. Results point to the better performance of MultiBoost in our experimental conditions. Also, the use in the training process of equal number of examples for all the classes and discarding frames from the initial parts of nasals vowels contributes to better performances.

Overall, the best results obtained are inside the interval reported in [5] for American English. But, contrary to [5], we didn't achieve good results with the SVM classifier.

Future research will include: experiments with other classifiers , features selection, adding a 3rd level of classification to individualize the vowels, experiments on classifiers output fusion and, our main goal, use of the detected frames in a phonetic-based speaker verification system for EP.

7. Aknowledgments

First author work was made possible by a scholarship from the Portuguese Science and Technology Foundation (PhD grant SFRH/BD/18230/2004).

8. References

[1] M. Sharma and R. J. Mammone, ""blind" speech segmentation: automatic segmentation of speech without linguistic knowledge," in *ICSLP*, 1996, pp. 1237–1240.

[2] A. Kanae, S. Tostomu, and A. Takayuki, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoust Sci Technol*, vol. 27, no. 4, pp. 233–235, 2006.

[3] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, pp. 225–239, 2004.

[4] T. Pruthi and C. Y. Espy-wilson, "Acoustic parameters for the automatic detection of vowel nasalization," in *INTERSPEECH 2007*, Antwer, Belgium, August 2007.

[5] T. Pruthi, "Analisys, vocal-tract modeling and automatic detection of vowel nasa-lization." Ph.D. dissertation, University of Marylans, UMI, 2007.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 195, ch. 14, pp. 495–516.

[7] D. G. Childers, *Speech processing and synthesis toolboxes*. John Wiley & Sons, Inc., 2000.

[8] D. A. Cairns, J. H. L. Hansen, and J. E. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 35–45, 1996.

[9] M. K. Huffman, "The role of F1 amplitude in producing nasal percepts," *Journal of the Acoustical Society of America*, vol. 88, no. S1, 1990.

[10] J. Glass, "Nasal consonants and nasalized vowels an acoustic study and recognition experiment," Master's thesis, MIT, Cambridge, USA, 1984.

[11] G. Coranin and M. Zaffalon, "Learning reliable classifiers from small or incomplete data sets: The Naive Credal Classifier 2," *Journal of Machine Learning Research*, vol. 9, pp. 581–621, 2008.

[12] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, 2005.

[13] G. Coranin and M. Zaffalon, "JNCC2: The java implementation of naive credal classifier 2," *Journal of Machine Learning Research*, vol. 9, pp. 2695–2698, 2008.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," 1995.