Can Ultrasonic Doppler Help Detecting Nasality for Silent Speech Interfaces?

An Exploratory Analysis based on Alignement of the Doppler Signal with Velum Aperture Information from Real-Time MRI

João Freitas[1,2], António Teixeira[2] and Miguel Sales Dias[1,3]
[1]*Microsoft Language Development Center, Lisboa, Portugal*
[2]*Dep. Electronics Telecommunications & Informatics/IEETA, University of Aveiro, Portugal*
[3]*ISCTE-Lisbon University Institute/ADETTI-IUL, Lisboa, Portugal*
jdcfreitas@live.com.pt, ajst@ua.pt, midias@microsoft.com

Keywords: Nasality Detection for Silent Speech Interaction, Velum Movement Detection, Ultrasonic Doppler, Nasal Vowels, Portuguese.

Abstract: This paper describes an exploratory analysis on the usefulness of the information made available from Ultrasonic Doppler signal data collected from a single speaker, to detect velum movement associated to European Portuguese nasal vowels. This is directly related to the unsolved problem of detecting nasality in silent speech interfaces. The applied procedure uses Real-Time Magnetic Resonance Imaging (RT-MRI), collected from the same speaker providing a method to interpret the reflected ultrasonic data. By ensuring compatible scenario conditions and proper time alignment between the Ultrasonic Doppler signal data and the RT-MRI data, we are able to accurately estimate the time when the velum moves and the type of movement under a nasal vowel occurrence. The combination of these two sources revealed a moderate relation between the average energy of frequency bands around the carrier, indicating a probable presence of velum information in the Ultrasonic Doppler signal.

1 INTRODUCTION

A known challenge in Silent Speech Interfaces (SSI), including those based on Ultrasonic Doppler Sensing (UDS) (Freitas *et al*., 2012a), is the detection of the nasality phenomena in speech production, being unclear if information on nasality is present in the UDS signal. Nasality is an important characteristic of several languages, such as French and European Portuguese (EP) (Teixeira, 2000), being the latter the selected language for the experiments here reported. Additionally, it has been shown before, that nasality can cause severe word recognition degradation in UDS (Freitas *et al*., 2012a) and Surface Electromyography (Freitas *et al*., 2012b) based interfaces for this language.

An SSI can be seen as a possible alternative to conventional speech interfaces since they allow for communication to occur in the absence of an acoustic signal. It brings advantages when used in situations where privacy or confidentiality is required, in the presence of environmental noise, such as in office settings, or when used by speech-impaired persons such as those who were subjected to a laryngectomy, making it a suitable candidate for an interface to be used in Ambient Assisted Living scenarios. An UDS-based SSI could eventually be included in a multimodal interface as one of the core input modalities (Zhu *et al*., 2007, Freitas *et al*., 2013).

The UDS approach main advantages are: its non-invasive nature, since the device is completely non-obtrusive and it has been proven to work without requiring any attachments; not being affect by environment noise in the audible frequency range; the required hardware is commercially available; and is very inexpensive. These advantages make UDS an interesting approach and an attractive research topic in the area of Human-Computer Interaction (HCI) (Raj *et al*., 2012). The sensing method is based on the emission of a pure tone in the ultrasonic range towards the moving target and the reflected signal is captured by an ultrasound receiver tuned to the transmitted frequency. The movement

of the target will cause Doppler shifts in the reflected signal, creating components at different frequencies, proportional to their velocity relative to the sensor. This technique has been applied to many areas of speech technology (Kalgaonkar and Raj, 2008; Toth *et al*., 2010), including speech and silent speech recognition (Srinivasan *et al*., 2010; Freitas *et al*., 2012a).

This paper describes an exploratory analysis on the existence of velum movement information detected in the Ultrasonic Doppler signal. The reflected signal contains information about the articulators and the moving parts of the face of the speaker, however, it is yet unclear how to distinguish between articulators and if velum movement information is actually being captured. Therefore, considering our aim of detecting velum movement and to provide a ground truth for our research, we used images collected from Real-Time Magnetic Resonance Imaging (RT-MRI) and extracted the velum aperture information during the nasal vowels of European Portuguese. Then, by combining and registering these two sources, ensuring compatible scenario conditions and proper time alignment, we are able to accurately estimate the time when the velum moves and the type of movement (i.e. ascending or descending) under a nasal vowel production phenomenon. Using this method we are able to correlate the features extracted from the UDS signal with the signal that represents the velum movement and analyse if velum information is being captured in our UDS signal analysis, for all nasal vowels.

The remainder of this paper is structured as follows: section 2 presents background notions of the nasality phenomenon and its impact on European Portuguese, as well as a description of how the Doppler Effect works; section 3 presents UDS related work in the area of HCI; section 4 describes the methodology used for extracting information from the RT-MRI images, the UDS device, how both signals were synchronized and the features extracted from the Ultrasonic signal; in section 5 the results of our exploratory analysis are presented; in section 6 we discuss these results and finally, in section 7, we present the conclusions of this study.

2 BACKGROUND

2.1 Nasality in European Portuguese

The production of a nasal sound involves air flow through the oral and nasal cavities. This air passage for the nasal cavity is essentially controlled by the velum that when lowered allows for the velopharyngeal port to be open, enabling resonance in the nasal cavity and the sound to be perceived as nasal. The production of oral sounds occurs when the velum is raised and the access to the nasal cavity is closed (Teixeira, 2000).

Nasality is a common characteristic of several languages around the world, however, only 20% of these languages have nasal vowels (Rossato *et al*. 2006). In EP there are five nasal vowels ([ɐ̃, ẽ, ĩ, õ, ũ])); three nasal consonants ([m], [n], and [ɲ]); and several nasal diphthongs [wɐ̃] (e.g. *quando*), [wẽ] (e.g. *aguentar*), [jɐ̃] (e.g. *fiando*), [wĩ] (e.g. *ruim*) and triphthongs [wɐ̃w] (e.g. *enxaguam*). Nasal vowels also diverge among languages, for example, nasal vowels in EP differ from French in its wider variation in the initial segment and stronger nasality at the end (Trigo, 1993; Lacerda and Head, 1966). Differences at the pharyngeal cavity level and velum port opening quotient were also detected by Martins *et al*. (2008) when comparing the articulation of EP and French nasal vowels.

2.2 The Doppler Effect

The Doppler Effect is the modification of the frequency of a wave when the observer and the wave source are in relative motion. If v_s and v_o are the speed of the source and the observer measured on the direction observer-source, if *c* is the propagation velocity of the wave on the medium and if f_0 is the source frequency, the observed frequency will be:

$$f = \frac{c + v_o}{c + v_s} f_0 \quad (1)$$

Considering a standstill observer $v_o = 0$ and $v_s \ll c$ the following approximation is valid:

$$f = \left(1 - \frac{v_s}{c}\right) f_0 \text{ or } \Delta f = -\frac{v_s}{c} f_0 \quad (2)$$

We are interested in echo ultrasound to characterize the moving articulators of a human speaker. In this case a moving body with a speed v (positive when the object is moving towards the emitter/receiver) reflects an ultrasound wave, whose frequency is measured by a receiver placed closely to the emitter. The observed Doppler shift will then be the double:

$$\Delta f = \frac{2v}{c} f_0 \quad (3)$$

3 RELATED WORK

In this section we present the work related with Ultrasonic sensors applied to HCI, in particular to speech recognition. Ultrasonic sensors have been applied to diverse and multiple areas that go from industrial automation to medical solutions, however, only in 1995 this technology was applied to speech recognition by Jennings and Ruck (1995), presenting the first "Ultrasonic Mike" with the goal of improving automatic speech recognition in noisy environments. In their work, Jennings and Ruck used an emitter and a receiver based on piezoelectric material and a 40 kHz oscillator to create a continuous wave ultrasonic signal.

More than a decade later, in 2007, Ultrasonic Doppler research saw new developments being applied to distinct areas of HCI, including speech recognition (Zhu *et al*., 2007). Since then, Ultrasonic Doppler has been applied to characterization and analysis of human gait (Kalgaonkar and Raj, 2007), gesture recognition (Kalgaonkar and Raj, 2009), speaker recognition (Kalgaonkar and Raj, 2008), speech synthesis (Toth *et al*., 2010), voice activity detection (Kalgaonkar *et al*., 2007), silent speech (Freitas *et al*. 2013) and speech recognition (Srinivasan *et al*., 2010; Freitas *et al*., 2012).

Still, several issues that can be found in the state-of-the-art remain unsolved: speaker dependence, sensor distance sensitivity, spurious movements made by the speaker, silent articulation, amongst others. Since Doppler shifts capture the articulators' movement, we believe that some of these problems can be attenuated or even solved if information about each articulator can be extracted.

In terms of UDS signal analysis Livescu *et al*. (2009) studied the phonetic discrimination in the UDS signal. In this study the authors tried to determine a set of natural sub-word units, concluding that the most prominent groupings of consonants include both place and manner of articulation classes and, for vowels, the most salient groups include close, open and round vowels.

In this paper we focus on determining if a particular articulator – the velum – is actually captured by the sensor and determine in which cases it is more evident by looking at the occurrence of nasal vowels in EP, a language with strong and particular nasal characteristics.

4 DATA COLLECTION, SYNCHRONIZATION AND FEATURE EXTRACTION

In order to understand if velum movement information can be found in the Doppler shifts of the echo signal, a signal that describes the velum movement is used as a reference. This signal was extracted from RT-MRI images, as described in section 4.2. This section also describes the hardware and setup of the Ultrasonic device and how synchronization of both signals is achieved.

4.1 Ultrasonic Doppler Setup

A custom build device, depicted on Figure 1, with a dedicated circuit board was developed based on the work of Zhu (2008). It includes 1) the ultrasound transducers (400ST and 400SR working at 40 kHz) and a microphone to receive the speech signal; 2) a crystal oscillator at 7.2 MHz and frequency dividers to obtain 40 and 36 kHz; 3) all the amplifiers and linear filters needed to process the echo signal and the speech signal. Since the board is placed in front of the speaker, the echo signal will be the sum of the contributions of all the articulators. If the ultrasound generated is a sine wave $sin2\pi f_0\, t$, an articulator with a velocity v_i will generate an echo wave that can be characterized by:

$$x_{\mathrm{i}} = a_{\mathrm{i}} \sin 2\pi f_0 \left(t + \frac{2}{c} \int_0^t v_{\mathrm{i}}\, \mathrm{d}\tau + \varphi_{\mathrm{i}} \right) \qquad (4)$$

$a_{\mathrm{i}}, \varphi_{\mathrm{i}}$ are parameters defining the reflection and are function of the distance. Although they are also function of time they are slow varying and are going to be considered constants. The total signals will be the sum for all articulators and the moving parts of the face of the speaker

$$x = \sum\nolimits_{\mathrm{i}} a_{\mathrm{i}} \sin 2\pi f_0 \left(\mathrm{t} + \frac{2}{\mathrm{c}} \int_0^{\mathrm{t}} v_{\mathrm{i}}\, \mathrm{d}\tau + \varphi_{\mathrm{i}} \right) \qquad (5)$$

The signal is a sum of frequency modulated signals. It was decided to make a frequency translation by multiplying the echo signal by a sine wave of a frequency $f_a = 36\ kHz$ and low passing the result it is obtained a similar frequency modulated signal centered at $f_1 = f_0 - f_a$, i.e., $f_1 = 4\ kHz$.

$$d = \sum\nolimits_{\mathrm{i}} a_{\mathrm{i}} \sin 2\pi \mathrm{f}_1 \left(t + \frac{2}{c} \int_0^{\mathrm{t}} v_{\mathrm{i}}\, \mathrm{d}\tau + \varphi_{\mathrm{i}} \right) \qquad (6)$$

This analogue operation is performed on the board and it was used an analogue multiplier AD633. The

Doppler echo signal and speech are then digitized at 44.1 kHz and the following process is digital and implemented in Matlab.

Figure 1: Custom built UDS device with two ultrasound transducers and a microphone.

4.2 RT-MRI Data Collection

The RT-MRI data collection was previously conducted at IBILI/Coimbra for nasal production studies. Images were acquired at the mid-sagittal and coronal oblique planes of the vocal tract (see Figure 2) using an Ultra-Fast RF-spoiled Gradient Echo (GE) pulse sequence and yielding a frame rate of 14 frames/second. Each recorded sequence contained 75 images. Additional information concerning the image acquisition protocol can be found in Silva *et al*. (2012).

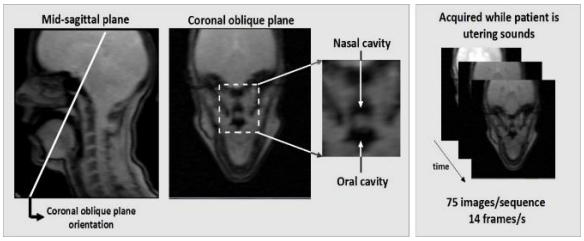

Figure 2: From left to right: mid-sagittal plane depicting orientation of the oblique plane used during acquisition, sample oblique plane showing the oral and nasal cavities and image sequence details (Teixeira *et al*., 2012).

Audio was recorded simultaneously with the real-time images, inside the scanner, at a sampling rate of 16 kHz, using a fiber optic microphone. For synchronization purposes a TTL pulse was generated from the RT-MRI scanner (Teixeira et al. 2012).

4.3 Extraction of information on nasal port from RT-MRI data

For the mid-sagittal RT-MRI sequences of the vocal tract, since the main interest was to interpret velum position/movement from the sagittal RT-MRI sequences, instead of measuring distances (e.g., from velum tip to the posterior pharyngeal wall), we opted for a method based on the area variation between the velum and pharynx, closely related to velum position.

An image with the velum fully lowered was used to define a region of interest (ROI). Then, a region growing algorithm was applied with a seed defined in a hypo intense pixel inside the ROI. This ROI is roughly positioned between the open velum and the back of the vocal tract and the main purpose is that the velum will move over that region when closing. Since this first ROI could be defined enclosing also a larger region, even including a part of the velum (which will not influence the process), it is only important that the seed is placed in a dark (hypo intense) pixel inside it, in order to exclude the most of the velum from the region growing when it is positioned inside the ROI. Figure 3 presents the contours of the segmented region over different image frames encompassing velum lowering and rising. For representation purposes, in order not to occlude the image beneath, only the contour of the segmented region is presented. Processing is always performed over the pixels enclosed in the depicted region. Notice that the white boundaries presented in the images depict the result of the region growing inside the defined ROI (which just limits the growth) and not the ROI itself. The number of hypo intense pixels (corresponding to an area) inside the ROI decreases when the velum closes and increases when the velum opens. Therefore, a closed velum corresponds to area minima while an open velum corresponds to local area maxima, which allows detecting the frames where the velum is open. Since for all image sequences there was no informant movement, the ROI has only to be set once, for each informant, and can then be reused throughout all the processed sagittal real-time sequences. After ROI definition (around one minute and reusable throughout all image sequences from the same speaker), setting a seed, revising the results and storing the data took one minute per image sequence.

These images allowed deriving a signal over time that describes the velum movement (also shown in Figure 3 and depicted as dashed line in Figure 4). As can be observed, minima correspond to a closed velopharingeal port (oral sound) and maxima to an open port (nasal sound).

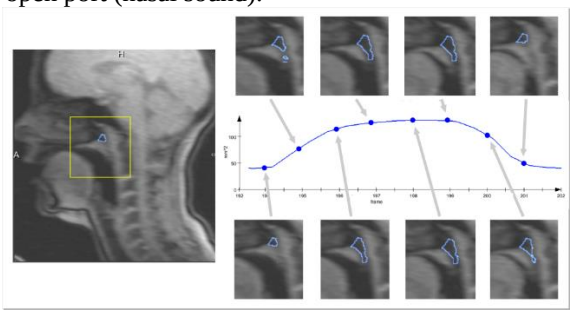

Figure 3: Mid-sagittal RT-MRI images of the vocal tract for several velum positions, over time, showing evolution from a raised velum, to a lowered velum and back to initial conditions. The presented curve, used for analysis, was derived from the images.

4.4 Corpora

The corpora used in this study, both RT-MRI and UDS, share a set of prompts composed by several non-sense words that contain five EP nasal vowels ([ɐ̃, ɛ̃, ĩ, õ, ũ]) isolated and in word-initial, word-internal and word-final context (e.g. ampa [ɐ̃pɐ], pampa [pɐ̃pɐ], pam [pɐ̃]). The nasal vowels are flanked by the bilabial stop or the labiodental fricative. This set contains 3 utterances per nasal vowels and data from a single speaker. The UDS data was recorded at a distance of 12 cm from the speaker.

4.5 Signals synchronization

In order to be able to take advantage of the RT-MRI velum information we need to synchronize the UDS and RT-MRI signals. We start by aligning both UDS and the information extracted from the RT-MRI with the corresponding audio recordings. We resample the audio recordings to 12 kHz and apply Dynamic Time Warping (DTW) to the signals, finding the optimal match between the two sequences. Based on the DTW result we map the information extracted from RT-MRI from the original production to the UDS time axis, establishing the needed correspondence between the UDS and the RT-MRI information, as depicted on Figure 4.

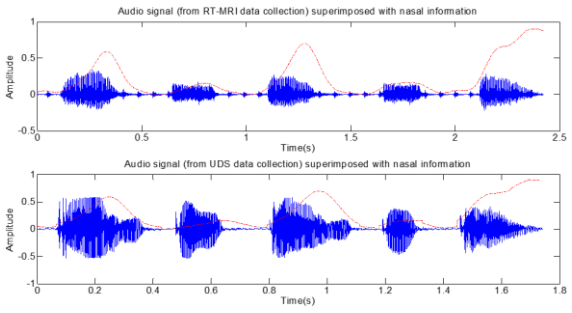

Figure 4: Exemplification of the warped signal representing the nasal information extracted from RT-MRI (dashed line) superimposed on the speech recorded during the corresponding RT-MRI and UDS acquisition, for the sentence [ɐ̃pɐ, pɐ̃pɐ, pɐ̃].

4.6 UDS Feature Extraction

For this experiment we have selected two types of features - frequency-band energy averages and energy-band frequency averages (Livescu *et al.*, 2009; Zhu, 2008). To obtain the frequency-band energy averages, we split the signal spectrum into several non-linearly divided bands centered around the carrier. Then, the mean energy is computed for each band. The frequency interval for each band n is given by:

$$Interval_n = [fmin_n, fmax_n], -5 \leq n \leq 4 \quad (7)$$

where $fmin_0 = 4000\,Hz$ (carrier frequency), $fmin_n = fmax_{n-1}$, $fmax_n = fmin_n + \alpha\,(|n| + 1)$, and $\alpha = 40\,Hz$. As such, the bandwidths slowly increase from 40 Hz to 280 Hz, capturing higher resolution information near the carrier.

In order to compute the energy-band frequency averages we split the spectrum into several energy bands and compute frequency centroid for each band. We extract values from 14 bands (7 below and 7 above the carrier frequency) using 10 dB energy thresholds that range from 0 dB to -70 dB.

5 EXPLORATORY ANALYSIS

In order to achieve our aim of finding if velum movement information is present in the ultrasonic signal, we decided to measure the strength of association between the obtained features, which describes the ultrasonic signal and RT-MRI information and is an accurate representation of the ground truth. Below, we present several results based on Pearson's product-moment correlation coefficient, which measures how well the two signals are related and also the results of Independent Component Analysis application to the extracted features. The correlation values range between -1 and 1, thus the greater the absolute value of a correlation coefficient, the stronger the linear relationship is. The weakest relationship is indicated by a correlation coefficient equal to 0.

5.1 Results

When comparing the RT-MRI velum information with the obtained features along each frequency band, based on correlation magnitude presented in Figure 5, it is not clear which band presents the higher correlation, although the values near the

carrier are slightly higher. However, if we split our analysis by vowel, more interesting results are visible. Figure 6 shows the correlation results for utterances where only the nasal vowel [ɐ̃] occurs (e.g. ampa [ɐ̃pɐ], pampa [pɐ̃pɐ], pan [pɐ̃]) and it is visible a more distinct group of correlation values at the frequency interval [4040..4120] Hz. When looking at the nasal vowel [ẽ] a stronger correlation is also noticed in that interval. However, in the case of the nasal vowel [õ] and [ũ] higher correlation values are found in the [3880..4040] Hz range, with an average correlation magnitude of 0.42 for [õ] and 0.44 for [ũ] (depicted in Figure 7). For the nasal vowel [ĩ], we find much lower correlation values when compared with the remaining vowels such as [ɐ̃], [õ] or [ũ] and the best interval can be found in the [4240..4400] Hz range with an average correlation magnitude of 0.25.

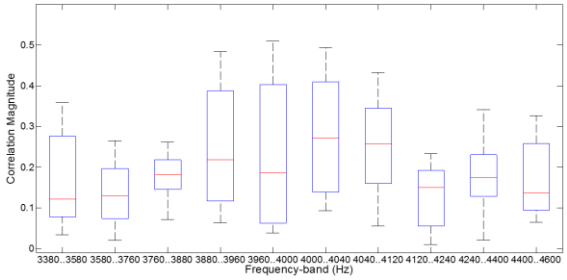

Figure 5: Boxplot for all utterances. The x axis lists the frequency-band features and the y axis corresponds to the absolute Pearson's correlation value. The central mark is the median and the edges of the box are the 25th and 75th percentiles.

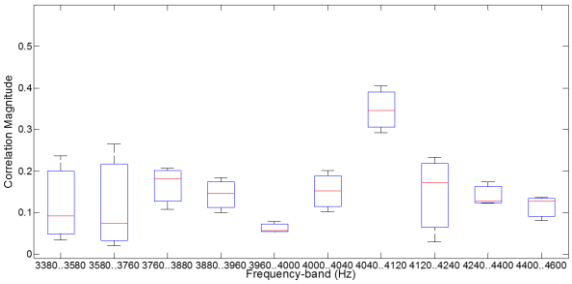

Figure 6: Boxplot for utterances with [ɐ̃]. The x axis lists the frequency-band features and the y axis corresponds to the absolute Pearson's correlation value.

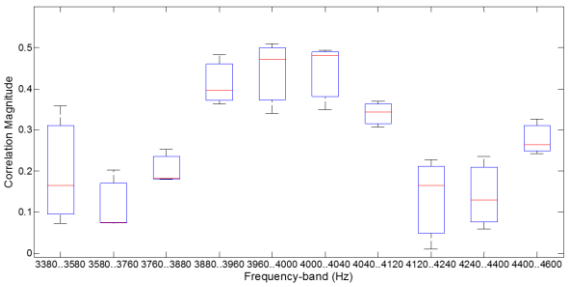

Figure 7: Boxplot for utterances with [ũ]. The x axis lists the frequency-band features and the y axis corresponds to the absolute Pearson's correlation value.

When looking at the energy-band features for all vowels we find similar values for the energy bands below -30dB, where the highest average correlation value is achieved by the [-30..-40] dB range above and below the carrier with 0.23. If we split out analysis by vowel, the highest value is achieved by the nasal vowel [õ] with an average correlation of 0.43 for the [-40..-50] dB interval above the carrier. The second best result using energy-band features is obtained by the nasal vowel [ũ] in the [-30..-40] dB range with values of 0.40 above the carrier and 0.39 below the carrier.

5.1.1 Applying Independent Component Analysis

As mentioned earlier the Ultrasonic Doppler signal can be seen as the sum for all articulators and the moving parts of the face of the speaker. Thus, the signal can be interpreted as a mix of multiple signals. Considering our goal, an ideal solution would be to find a process to isolate the signal created by the velum. Independent Component Analysis (ICA) is a method used for separating a multivariate signal with independent sources linearly mixed, thus the underlying idea is to understand if by applying blind source separation we can obtain independent components that relate with each articulator movement, including the velum.

For that purpose we applied the FastICA algorithm (Hyvarinen, 1999) using the RT-MRI information as *a priori* to build the separating matrix. This allowed to obtain independent components with a higher correlation value than when compared to the extracted features without any transformation, as shown in Table 1. Also, due to the singularity of the covariance matrix we observe a dimensionality reduction of 4 to 8 components

depending on the utterance when using frequency-band features. When using energy-band features we observe a dimensionality reduction of 6 to 12 components.

Table 1: Average correlation magnitude values with 95% confidence interval using frequency-band and energy band features for the best independent components of each utterance.

	Average correlation magnitude	
	Frequency	*Energy*
All vowels	0.42 ± 0.05	0.41 ± 0.04
[ɐ̃]	0.44 ± 0.04	0.33 ± 0.05
[ẽ]	0.41 ± 0.09	0.41 ± 0.10
[ĩ]	0.30 ± 0.05	0.42 ± 0.03
[õ]	0.47 ± 0.08	0.41 ± 0.05
[ũ]	0.48 ± 0.14	0.50 ± 0.07

6 DISCUSSION

The applied methodology uses the audio signal to synchronize two distinct signals that otherwise were very hard to align. Although the two sources of information were recorded at different times, it is our belief that by reproducing the articulation movements we are able to obtain a very good indication of how the velum behaves upon the same stimulus for most cases. The utterances containing the [ũ] nasal vowels presented some alignment inaccuracies mainly at the end of the first phoneme and further improvements need to be considered for this particular case.

Knowing that the velum is a slow articulator, as shown by the RT-MRI velum movement information in Figure 4, and considering equation 2, it is expected that velum movement, if detected by UDS, is found in the regions near the carrier, which is where the results for [ɐ̃, ẽ, õ and ũ] present higher correlation. However, the velum is not the only slowly moving articulator and a different corpora which allows, for example, to discard jaw movements should be considered for future studies.

Another point of discussion is the differences found between nasal vowels. When looking at the correlation results of frequency-band features, a difference is noticed from [ĩ] to the remaining vowels. One possible explanation for this difference might be the articulation variances of each nasal vowel previously reported in literature (Schwartz, 1968). Since our technique is based on the reflection of the signal it is plausible that the tongue position influences the detection of the velum, particularly for the case of [ĩ] in which the tongue posture may block the UDS signal.

It would also be expected to find a clear difference between close and open vowels (Livescu et al., 2009). Although this is true for the nasal vowel [ĩ], it was not verified in the [ũ] case, which presented the highest correlation values along with [õ]. Further investigation is required to understand if for example the rounding of the lips during the articulation of these two vowels is influencing the signal reflection and in which way.

In this study we have also applied blind source separation as an attempt to split the signal into independent components. This technique has given slightly better results for both sets of features, showing that isolating the velum movement in the Doppler shifts might be possible. It is also noteworthy the fact that this process has led to a dimensionality reduction of 4 to 8 components depending on the utterance, which may have a relation with the number of mobile articulators that can cause Doppler shifts in the signal (i.e. tongue, lower jaw, velum, lips, cheeks, oral cavity).

7 CONCLUSIONS

This paper analysis the presence of information about the velum movement for European Portuguese nasal vowels in the Ultrasonic Doppler signal. As ground truth for our study, we use previously collected RT-MRI information from the same speaker and, after extracting a signal that describes the movement of the velum, we apply a synchronization technique based on the audio signal collected from both *corpora*. With this approach we are able to estimate the velum behaviour and measure the strength of association between the features that describe the ultrasonic signal data and RT-MRI data, via computing the Pearson's product-moment correlation coefficient.

The obtained results show that for features based on the energy of pre-determined frequency bands, we are able find moderate correlation values, for the case of the vowels [ɐ̃], [õ] and [ũ] and weaker correlation values in the [ĩ] case. Moderate correlation values were also found using energy based features for bands below -30dB. We have also applied a blind source separation technique obtaining components with a better description of the velum movement.

For future work and based on this methodology, we plan to apply the same process to other

articulators such as the tongue or lips, which will help to determine important aspects and more details about the captured information. We also intend to expand the current *corpora* with more speakers and adequate prompts for these scenarios. It would also be important to analyse the impact of distance from the UDS emitter to the speaker face in the captured information.

ACKNOWLEDGEMENTS

This work was partially funded by Marie Curie Golem (ref.251415, FP7-PEOPLE-2009-IAPP), by FEDER through the Program COMPETE under the scope of QREN 5329 FalaGlobal (PTDC/EEA-PLP/098298/2008) and by National Funds (FCT- Foundation for Science and Technology) in the context of IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEst-C/EEI/UI0127/2011) and in the context of the Project HERON II (PTDC/EEA-PLP/098298/2008). The authors would also like to thank the speaker involved in the experiment.

REFERENCES

Freitas, J., Teixeira, A., Dias, M. S., 2012b. Towards a Silent Speech Interface for Portuguese: Surface Electromyography and the nasality challenge. In *Int. Conf. on Bio-inspired Systems and Signal Processing*, Vilamoura, Algarve, Portugal.

Freitas, J., Teixeira, A., Dias, M. S., 2013. Multimodal silent speech interface based on Video, depth, surface electromyography, and ultrasonic Doppler, Workshop on Speech Production in Automatic Speech Recognition, Lyon, France.

Freitas, J., Teixeira, A., Vaz, F., Dias, M. S., 2012a. Automatic Speech Recognition based on Ultrasonic Doppler Sensing for European Portuguese. *Advances in Speech and Language Technologies for Iberian Languages*, vol. CCIS 328, Springer.

Hyvarinen, A, 1999. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on* Vol. 10, no. 3, pp. 626-634.

Jennings, D. L., Ruck, D. W, 1995. Enhancing automatic speech recognition with an ultrasonic lipmotion detector, In *Int. Conf. on Acoustics, Speech, and Signal Processing*, Detroit.

Kalgaonkar, K. and Raj, B., 2007. Acoustic Doppler Sonar for Gait Recognition, *IEEE International Conference on Advance Video and Signal-based Surveillance (AVSS2007)*.

Kalgaonkar, K. and Raj, B., 2008. Ultrasonic Doppler Sensor for Speaker Recognition, *IEEE Intl. Conf. on Acoustics Speech and Signal Processing 2008.*

Kalgaonkar, K. and Raj, B., 2009. One-handed Gesture Recognition using Ultrasonic Doppler Sonar, *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing 2009*.

Kalgaonkar, K., Raj B., Hu., R., 2007. Ultrasonic doppler for voice activity detection. *IEEE Signal Processing Letters*, vol.14(10), pp. 754–757.

Lacerda, A. and Head, B. F., 1996. Análise de sons nasais e sons nasalizados do Português. *Revista do Laboratório de Fonética Experimental* (de Coimbra). VI:5_70.

Livescu, K., Zhu, B. and Glass, J., 2009. On the phonetic information in ultrasonic microphone signals" *Intl. Conf. on Acoustics, Speech and Signal Processing 2009*.

Martins, P. Carbone, I. Pinto, A. Silva, A. and Teixeira, A., 2008. European Portuguese MRI based speech production studies. *Speech Communication*. NL: Elsevier, Vol.50, No.11/12, ISSN 0167-6393, pp. 925–952.

Raj, B., Kalgaonkar, K. Harrison, C. and Dietz, P., 2012. Ultrasonic Doppler Sensing in HCI. *Pervasive Computing, IEEE 11, no. 2, pp. 24-29*.

Rossato, S. Teixeira, A. and Ferreira, L., 2006. Les Nasales du Portugais et du Français: une étude comparative sur les données EMMA. *In XXVI Journées d'Études de la Parole*. Dinard, France.

Schwartz, M. F. 1968. The acoustics of normal and nasal vowel production. *Cleft Palate Journal* 5: 125-40.

Silva, S., Martins, P., Oliveira, C., Silva, A., Teixeira, A., 2012. Segmentation and Analysis of the Oral and Nasal Cavities from MR Time Sequences. Image Analysis and Recognition. *Proceedings of ICIAR 2012*, LNCS, Springer.

Srinivasan, S., Raj, B., Ezzat., T., 2010. Ultrasonic sensing for robust speech recognition, In *Intl. Conf. on Acoustics, Speech, and Signal Processing 2010.*

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., Shosted, R., 2012. Real-time MRI for Portuguese: database, methods and applications, *Proceedings of PROPOR 2012*, LNCS vol. 7243. pp. 306-317.

Teixeira, J. S., 2000. Síntese Articulatória das Vogais Nasais do Português Europeu. *PhD Thesis*, Universidade de Aveiro.

Toth, A.R., Kalgaonkar, K., Raj, B., T. Ezzat, 2010. Synthesizing speech from Doppler signals. *IEEE International Conference on Acoustics Speech and Signal Processing*, pp.4638-4641.

Trigo, R. L., 1993. The inherent structure of nasal segments, In *Nasals, Nasalization, and the Velum, Phonetics and Phonology*, M. K. Huffman e R. A. Krakow (eds.), Vol. 5, pp.369-400, Academic Press Inc.

Zhu, B., 2008. Multimodal speech recognition with ultrasonic sensors. Master's thesis. Massachusetts Institute of Technology, Cambridge, Massachusetts.

Zhu, B., Hazen, T. and Glass, J. R., 2007. Multimodal speech recognition with ultrasonic sensors. In *Eurospeech 2007*.