TOWARDS A SILENT SPEECH INTERFACE FOR PORTUGUESE

Surface Electromyography and the nasality challenge

João Freitas[1,2,3], António Teixeira[3] and Miguel Sales Dias[1,2]
[1]*Microsoft Language Development Center, Tagus Park, Porto Salvo, Portugal*
[2]*ISCTE-Lisbon University Institute/ADETTI-IUL, Lisboa,Portugal*
[3]*Departamento de Electrónica Telecomunicações e Informática/IEETA, Universidade de Aveiro, Portugal*
i-joaof@microsoft.com, ajst@ua.pt, midias@microsoft.com

Keywords: Silent Speech, Human-Computer Interface, European Portuguese, Surface Electromyography, Nasality

Abstract: A Silent Speech Interface (SSI) aims at performing Automatic Speech Recognition (ASR) in the absence of an intelligible acoustic signal. It can be used as a human-computer interaction modality in high-background-noise environments, such as living rooms, or in aiding speech-impaired individuals, increasing in prevalence with ageing. If this interaction modality is made available for users own native language, with adequate performance, and since it does not rely on acoustic information, it will be less susceptible to problems related to environmental noise, privacy, information disclosure and exclusion of speech impaired persons. To contribute to the existence of this promising modality for Portuguese, for which no SSI implementation is known, we are exploring and evaluating the potential of state-of-the-art approaches. One of the major challenges we face in SSI for European Portuguese is recognition of nasality, a core characteristic of this language Phonetics and Phonology. In this paper a silent speech recognition experiment based on Surface Electromyography is presented. Results confirmed recognition problems between minimal pairs of words that only differ on nasality of one of the phones, causing 50% of the total error and evidencing accuracy performance degradation, which correlates well with the exiting knowledge.

1 INTRODUCTION

Since the dawn of mankind, speech communication has been and still is the dominant mode of human communication and information exchange and, for this reason, spoken language technology has suffered considerable evolution in the last years, in the scientific community. However, conventional automatic speech recognition (ASR) systems mostly use a single source of information – the audio signal. When this audio signal becomes corrupted in the presence of environmental noise or assumes unexpected patterns, like the ones verified in elderly speech (Wilpon and Jacobsen, 1996), speech recognition performance degrades, leading users to opt for a different modality or give up using the system at all. These types of systems have also revealed to be inadequate for users without the ability to create an audible acoustic signal because of speech impairments (e.g. laryngectomy) or in situations where privacy or non-disturbance is required (Denby et al., 2010). In public environments where silence is often necessary such as, talks, cinema or libraries, someone talking is usually considered annoying, thus providing the ability to communicate or execute commands in these situations has become a point of common interest. Likewise, disclosure of private conversations can occur by performing a phone call in public places, which may lead to embarrassing situations from the caller point of view or even regarding information leaks.

To tackle these problems in the context of ASR for Human-Computer Interaction (HCI), a Silent Speech Interface (SSI) in European Portuguese (EP) is envisioned. By acquiring sensor data from elements of the human speech production process – from glottal, muscles and articulators activity, their neural pathways or the central nervous system itself, an SSI produces an alternative digital representation of speech, which can be recognized and interpreted as data, synthesized directly or routed into a communications network. Informally, one can say that a SSI extends the human speech production model, with signal data sensed by ultrasonic waves,

computer vision or other sources. This provides a more natural approach than currently available speech pathology solutions like, electrolarynx, tracheo-oesophageal speech and cursor-based text-to-speech systems (Denby et al., 2010).

Currently, to our knowledge, no SSI system exists for European Portuguese, leaving, for example, speakers of this language with speech impairments unable to interact with HCI systems based on speech. Furthermore, no study or analysis has been made regarding the adoption of a new Romance language with distinctive characteristics to this kind of systems, and the specific problems that may arise from applying existing work to EP remains unknown. A particularly relevant characteristic of EP are the nasal sounds (Strevens, 1954), which are expected to be a challenge to several SSI techniques (Denby et al., 2010). The adoption of SSIs to a new language and the procedures involved constitute by itself an extension to the current scientific knowledge in this area. Using the techniques described in literature and adapting them to a new language will provide novel information towards language independence and language adoption techniques. Considering the particular nasal characteristics associated with EP, it is expected to see performance deterioration in terms of recognition rates and accuracy using existent approaches. If this occurs, the root cause of the system performance deterioration needs to be identified and new techniques based on that information, need to be thought, for example, by adding a sensor that is able to capture the missing information. This will allow concluding the particular aspects that influence language expansion, language independency and limitations of SSIs for the EP case.

To achieve our goals of developing a SSI for Portuguese we have previously selected a set of modalities (Surface Electromyography, Visual Speech Recognition and Ultrasonic Doppler sensing) (Freitas et al., 2011) based on their non-invasive characteristics, cost and technological availability. For this work we will focus on the SSI approach based on Surface Electromyography (sEMG), which has achieved promising results in last years, for languages such as English and Japanese. The sEMG modality collects myoelectric activity information of the neck and facial muscles generated before articulation of speech. This SSI modality is able to collect information from audibly uttered speech, murmurs or silent speech, being consequently robust do adverse environments such as public places, information disclosure and disturbance of bystanders.

The remainder of this document is structured as follows: In section 2, relevant background knowledge concerning the human speech production process is presented; Section 3 describes the related work and the state-of-the-art in EMG-based recognition; Section 4 describes and discusses the EMG-based recognition experiment in European Portuguese, including the observation and accuracy impact of the nasality phenomena; Finally, the conclusions and possible solutions for the appointed problems are presented in section 5.

2 BACKGROUND

The following section presents a brief description the speech production process, focusing on the muscles and articulators involved in this process. It also presents a brief description of the European Portuguese characteristics and related work for this language.

2.1 Speech motor control

Speech production requires a particularly coordinated sequence of events to take place, being considered as the most complex sequential motor task performed by humans (Seikel et al., 2010). After an intent or idea that we wish to express have been developed and coded into a language, we will map it into muscle movements. This means that the motor impulse received by the primary motor cortex is the result of several steps of planning and programming that already occurred in other parts of the brain such as, Broca's area, supplementary motor area and pre motor area. The motor neuron then sends the signal from the brain to the exterior body parts. When the nerve impulse reaches the neuromuscular junction, the neurotransmitter acetylcholine is released. When a certain threshold is hit, the sodium channels open up causing an ion exchange that propagates in both directions along the muscle fiber membranes. The depolarization process and ion movement generates an electromagnetic field in the area surrounding the muscle fibers, which is referred in literature as the myoelectric signal (De Luca, 1979). These electrical potential differences generated by the resistance of muscle fibers, leads to voltage patterns that occur in the region of the face and neck that when measured at the correspondent muscles, provide means to collect information about the resultant speech. This myoelectric activity occurs independently of the acoustic signal, i.e. occurs either the subject produces normal, silent or murmured speech. A detailed overview about the physiology of myoelectric signals and speech motor control can be seen in Seikel et al. (2010) and Gerdle et al. (1999).

2.1.1 Muscles and articulation

Articulation in phonetics describes how humans produce speech sounds and which speech organs are involved in this process. The articulators may be mobile or passive and the mobile articulators are usually positioned in relation to a passive articulator, through muscular action, to achieve different sounds. Mobile articulators are the tongue, lower jaw, velum, lips, cheeks, oral cavity (fauces and pharynx), larynx and the hyoid bone and the immobile articulators are the alveolar ridge of the upper jaw, the hard palate and teeth (Seikel et al., 2010).

Facial and neck muscles represent a vital role in positioning the articulators and in shaping the air stream into recognizable speech. Muscles related with lip movement, tongue and mandibular movement will then be the most influent in speech production. Below, the main muscles of the face and neck used in speech production are described (Hardcastle, 1976): *Orbicularis oris*: This muscle can be used for rounding, closing the lips and pulling the lips against the teeth or adducting the lips. Since its fibers run in several directions, many other muscles blend in with it; *Levator anguli oris*: This muscle is responsible for raising the upper corner of the mouth and may assist in closing the mouth by raising the lower lip for the closure phase in bilabial consonants; *Zygomaticus major*: This muscle is used to retract the angles of the mouth. It has influence in the labiodental fricatives and in the production of the [s] sound; *Platysma*: The *platysma* is responsible for aiding the *depressor anguli oris* muscle lowering the bottom corners of the lips. The *platysma* is the closest muscle to the surface in the neck area; *Tongue*: The tongue plays a fundamental role in speech articulation and is divided into intrinsic and extrinsic muscles. The intrinsic muscles (*Superior* and *Inferior Longitudinal*; *Transverse*) mostly interfere with the shape of the tongue, aiding in palatal and alveolar stops, in the production of the [s] sound by making the seal between the upper and lower teeth and in the articulation of back vowels and velar consonants. The extrinsic muscles (*Genioglossus*; *Hyoglossus*; *Styloglossus*; and *Palatoglossus*) are responsible for changing the position of the tongue in the mouth as well as shape and are important in the production of most of the sounds articulated in the front of the mouth, in the production of the vowels and velars, in the release of alveolar stop consonants, and contributes to the subtle adjustment of grooved fricatives; *Anterior Belly of the Digastric*: this is one of the muscles used to lower the mandible, to pull the hyoid bone and the tongue up and forward for alveolar and high frontal vowel articulations and raising pitch.

For a sound to be perceived as nasal the soft palate must be positioned in a way that the opening for the nasal cavity is larger than the airway opening for the oral cavity. This relation in the oral/nasal opening will enable resonance in the nasal cavity and consequently produce nasal sounds (Teixeira, 2000). The described movement of the soft palate is supported by the following muscles (Hardcastle, 1976): *Levator veli palatini*: This muscle main function is to elevate and retract the soft palate; *Superior pharyngeal constrictor*: This is a muscle of the pharynx and when it contracts it narrows the pharynx upper wall, also elevating the soft palate; *Tensor palatini*: This muscle tenses and spreads the soft palate when elevating; *Palatoglossus:* Along with gravity, previous muscles relaxation and the *Palatopharyngeous*, this muscle is responsible for the lowering of the soft palate.

2.2 European Portuguese characteristics

According to Strevens (1954), when one first hears EP, the characteristics that distinguish it from other Western Romance languages are: "the large amount of diphthongs, nasal vowels and nasal diphthongs, frequent alveolar and palatal fricatives and the dark diversity of the l-sound". Although, EP presents similarities in vocabulary and grammatical structure to Spanish, the pronunciation significantly differs. Regarding co-articulation, which is "the articulatory or acoustic influence of one segment or phone on another" (Magen, 1997), it is shown by Martins et al. (2008) that European Portuguese stops, are less resistant to co-articulatory effects than fricatives.

2.2.1 Nasality

Although nasality is present in a vast number of languages around the world, only 20% have nasal vowels (Rossato et al., 2006). In EP there are five nasal vowels ([ĩ], [ẽ], [ɐ̃], [õ], and [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 in EP diverge from other languages with such type of vowels, such as French, in its wider variation in the initial segment and stronger nasality at the end (Trigo, 1993). Doubts still remain regarding tongue positions and other articulators during nasals production in EP, namely, nasal vowels (Teixeira et al., 2003). Differences at the pharyngeal cavity level and velum port opening quotient were also detected by Martins et al. (2008) when comparing EP and French nasal vowels articulation. In EP, nasality can

distinguish consonants (e.g. the bilabial stop consonant [p] becomes [m]), creating minimal pairs such as [katu]/[matu] and vowels, in minimal pairs such as [titu]/[tĩtu].

2.3 SSIs for European Portuguese

The existing SSI research has been mainly developed for English, with some exceptions for French (Tran et al., 2009) and Japanese (Toda et al., 2009). As mentioned, there is no published work for European Portuguese in the area of SSIs, apart from an initial and recent contribution of the authors with an experiment of an SSI for EP, using Visual Speech Recognition and Ultrasonic Doppler sensing techniques (Freitas et al. 2011), which also tackles nasality, although there is previous research on related areas, such as the use of Electromagnetic Articulography (Rossato et al., 2006), Electroglotograph and MRI (Martins et al., 2008) for speech production studies, articulatory synthesis (Teixeira and Vaz, 2000) and multimodal interfaces involving speech (Dias et al., 2009). There are also several studies on lip reading systems for EP that aim at robust speech recognition based on audio and visual streams (Pêra et al., 2004; Sá et al., 2003). However, none of these addresses European Portuguese distinctive characteristics, such as nasality.

3 RELATED WORK

The process of recording and evaluating this electrical muscle activity is called Electromyography (EMG). Currently, there are two sensing techniques to measure EMG signals: invasive indwelling sensing and non-invasive sensing. The work here presented will focus on the second technique, which is when the myoelectric activity is measured by non-implanted electrodes. These are usually attached to the subject on some adhesive basis, which may obstruct movement, especially when placed on facial muscles. By measuring facial muscles, the surface EMG (sEMG) electrodes will measure the superposition of multiple fields (Gerdle et al. 1999) and for this reason the resulting EMG signal should not be attributed to a single muscle and should consider the muscle entanglement verified in this part of the human body. The sensor presence can also cause the subject to alter his/her behaviour, be distracted or restrained, subsequently altering the experiment result. The EMG signal is not affected by noisy environments, however differences may be found in the speech production process in the presence of noise (Junqua et al., 1999). Muscle activity may also change in the presence of physical apparatus, such as mouthpieces used by divers, medical conditions such as laryngectomies, and local body potentials or strong magnetic field interference (Jorgensen and Dusan, 2010).

Surface EMG-based speech recognition, overcomes some of the major limitations found on automatic speech recognition based on the acoustic signal such as: non-disturbance of bystanders, robustness in acoustically degraded environments, privacy during spoken conversations and constitutes an alternative for speech-handicapped subjects (Denby et al., 2010). This technology has also been used for solving communication in acoustically harsh environments, such as the cockpit of an aircraft (Chan et al., 2001) or when wearing a self-contained breathing apparatus or a hazmat suit (Betts et al., 2006).

3.1 State-of-the-art

Relevant results in this area were first reported in 2001 by Chan et al. (2001) where five channels of surface Ag-AgCl sensors were used to recognize ten English digits. In this study accuracy rates as high as 93% were achieved. The same author (Chan, 2003) was the first to combine conventional ASR with sEMG with the goal of robust speech recognition in the presence of environment noise. In 2003, Jorgensen et al. (2003) achieved an average accuracy rate of 92% for a vocabulary with six distinct English words, using a single pair of electrodes for non-audible speech. However, when increasing the vocabulary to eighteen vowel and twenty-three consonant phonemes in later studies (Jorgensen and Binsted, 2005) using the same technique the accuracy rate decreased to 33%. In this study problems in the alveolars pronunciation and subsequently recognition using non-audible speech were reported and several challenges identified such as, sensitivity to signal noise, electrode positioning, and physiological changes across users. In 2007, Jou et al. (2007) reported an average accuracy of 70.1% for a 101-word vocabulary in a speaker dependent scenario. In 2010, Schultz and Wand (2010) reported similar average accuracies using phonetic feature bundling for modelling coarticulation on the same vocabulary and an accuracy of 90% for the best-recognized speaker.

In the last year's several issues of EMG-based recognition have been addressed such as, investigating new modeling schemes towards continuous speech (Jou et al., 2007; Schultz and Wand, 2010), speaker adaptation (Maier-Hein et al., 2005; Wand and Schultz, 2009) and the usability of the capturing devices (Manabe et al., 2003; Manabe

and Zhang, 2004). Latest research in this area has been focused on the differences between audible and silent speech and how to decrease the impact of different speaking modes (Wand and Schultz, 2011a); the importance of acoustic feedback (Herff et al., 2011); EMG-based phone classification (Wand and Schultz, 2011b); and session-independent training methods (Wand and Schultz, 2011c).

4 FIRST EXPERIMENTS WITH A SEMG-BASED SSI FOR PORTUGUESE

In our research we have designed an experiment to analyse and explore the sEMG silent speech recognition applied to European Portuguese. In this section an important research question is addressed: "is the sEMG SSI approach for European portuguese capable of distinguishing nasal sounds from oral ones?". To address this research problem, we have designed two scenarios where at first, we try to recognize arbitrary Portuguese words in order to validate our system and, in a second scenario, we want to recognize/distinguish words differing only by the presence or absence of nasality in the phonetic domain.

4.1 Acquisition setup

The used acquisition system hardware from Plux (2011) consisted of 4 pairs of EMG surface electrodes connected to a device that communicates with a computer via Bluetooth. These electrodes measure the myoelectric activity using bipolar surface electrode configuration, thus the result will be the amplified difference between the pair of electrodes, using a reference electrode located in a place with low or none muscle activity. As depicted on Figure 1, the sensors were attached to the skin using adhesive surfaces and their position followed some recommendations from previous studies found in literature (Jou et al., 2006), considering also a 2cm spacing between the electrodes centre and some hardware restrictions regarding unipolar configurations. The 4 electrodes pairs and their corresponding muscles are presented on Table 1. A reference electrode was placed on the mastoid portion of the temporal bone.

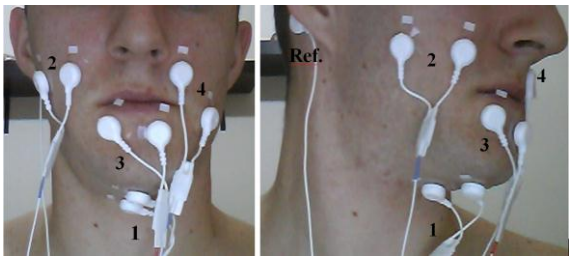

Figure 1: Surface EMG electrodes positioning.

Table 1: Electrode pair/muscle correspondence based on the configuration proposed by Jou et al. (2006).

Electrode pair	Muscle
1	Tongue and Anterior belly of the digastric
2	Zygomaticus major
3	Lower orbicularis oris
4	Levator angulis oris

The technical specifications of the acquisition system include sensors with a diameter of 10.0 mm and 3.95 mm of height, a voltage range that goes from 0.0V to 5.0V and a voltage gain of 1000.0. The recording signal was sampled at 600Hz and 12 bit samples were used.

In Figure 2 and Figure 3, observations of the raw EMG signal in the four channels for the minimal pair *cato/canto* are presented. Based on a subjective analysis we can see that the signals present similar temporal patterns where the tongue movement is first evidenced on channel 1 followed by the remaining muscles movement on the other channels.

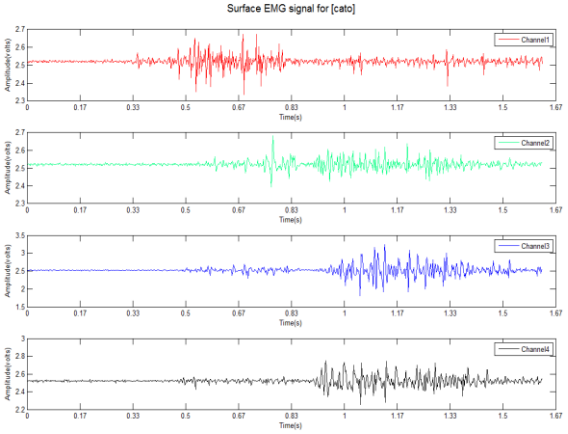

Figure 2: Surface EMG signal for the word *cato*.

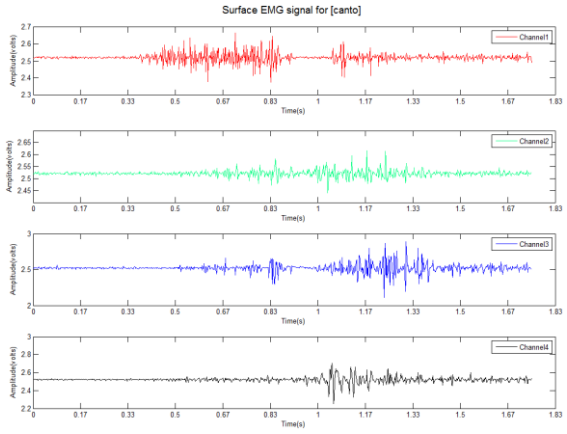

Figure 3: Surface EMG signal for the word *canto*.

4.2 Corpora

For this experiment two corpus - PT-EMG-A and PT-EMG-B – containing respectively 96 and 120 observation sequences like the ones depicted on Figure 2 and Figure 3, were created from scratch. All observations were recorded by a single speaker on a single recording session (no electrodes repositioning was considered). The PT-EMG-A consisted of 8 different European Portuguese words, 4 words that are part of a minimal pair where the presence or absence of nasality in one of its phones is the only difference and 4 digits, are described in Table 2. The PT-EMG-B corpus consisted also of 8 different words in European Portuguese, with 15 different observations of each word. However, for this corpus, the words represent four minimal pairs of words containing oral and nasal vowels (e.g. *cato*/*canto*) and sequences of nasal consonant followed by nasal or oral vowel (e.g. *mato*/*manto*). Table 3 lists the pairs of words used in the PT-EMG-B corpus and their respective phonetic transcription.

Table 2: Words that compose the PT-EMG-A corpus and their respective phonetic transcription.

Word List	Phonetic Transcription
Cato	[katu]
Peta	[petɐ]
Mato	[matu]
Tito	[titu]
Um	[ũ]
Dois	[dojʃ]
Três	[treʃ]
Quatro	[kwatru]

Table 3: Minimal pairs of words used in the PT-EMG-B corpus and their respective phonetic transcription.

Word Pair	Phonetic Transcription
Cato/ Canto	[katu] / [kɐ̃tu]
Peta / Penta	[petɐ] / [pẽtɐ]
Mato / Manto	[matu] / [mɐ̃tu]
Tito / Tinto	[titu] /[tĩtu]

4.3 Feature Extraction

For feature extraction we have used a similar approach to the one described by Jou (2006) based on temporal features instead of spectral ones or a combination of spectral plus temporal features, since it has been shown in previous studies (Jou et al., 2006) that time-domain features present better accuracy results. The extracted features are frame-based and for any given sEMG signal $s[n]$ frames of 30ms and a frame shift of 10ms is considered. Denoting $x[n]$ as the normalized mean of $s[n]$ and $w[n]$ as the nine-point double-averaged signal, a high-frequency signal $p[n]$ and $r[n]$ can be defined as:

$$r[n] = |p[n]| \tag{1}$$

$$p[n] = x[n] - w[n] \tag{2}$$

$$w[n] = \frac{1}{9}\sum_{n=-4}^{4} v[n] \tag{3}$$

$$v[n] = \frac{1}{9}\sum_{n=-4}^{4} x[n] \tag{4}$$

A feature f will then be defined as:

$$f = [\bar{w}, P_w, P_r, z_p, \bar{r}\,] \tag{5}$$

where $\bar{w}$, and $\bar{r}$ represent the frame-based time-domain mean, P_w and P_r the frame-based power, and z_p the frame-based zero-crossing rate as described below.

$$\bar{w} = \frac{1}{N}\sum_{n=0}^{N-1} w[n] \tag{6}$$

$$P_w = \frac{1}{N}\sum_{n=0}^{N-1} |w[n]|^2 \tag{7}$$

$$P_r = \frac{1}{N}\sum_{n=0}^{N-1}|r[n]|^2 \quad (8)$$

$$z_p = \text{zero-crossing of } p[n] \quad (9)$$

$$\bar{r} = \frac{1}{N}\sum_{n=0}^{N-1} r[n] \quad (10)$$

The feature vector also considers the concatenation of k adjacent frames as formulated below:

$$FV(f,k) = [f_{i-k}, f_{i-k+1}, \cdots, f_{i+k-1}, f_{i+k}] \quad (11)$$

where i is the current frame index. Most recent studies (Schultz and Wand, 2010) show that k =15 yields the best results.

In the end, the final feature vector is built by stacking the frame-based features of the four channels. In order to address the dimensionality of the resultant feature vector, PCA is applied reducing it to 32 coefficients per frame.

4.4 Classification

For this initial stage of research, the Dynamic Time Warping (DTW) classification technique was used to find an optimal match between the observations. DTW was chosen considering the relatively small number of observations and also because it addresses very well one of the characteristics of our problem: it provides temporal alignment to time varying signals that have different durations. This is precisely our case, since even observations of the pronunciation of the same word will certainly have different elapsed times.

In order to classify the results an algorithm, already applied to Visual Speech Recognition in Freitas et al. (2011), was used: (1) Randomly select K observations from each word in the selected corpus that will be used as the reference (training) pattern, while the remaining ones will be used for testing; (2) For each observation from the test group, compare the representative example and select the word that provides the minimum distance in the feature vector domain; (3) Compute WER, which is given by the number of incorrect classifications over the total number of observations considered for testing; (4) Repeat the procedure N times.

4.5 Results

Regarding the classification results for the corpus PT-EMG-A, the achieved values, using 20 iterations and K varying from 1 to 11, are listed on Table 4.

Table 4: Surface EMG WER classification results for the PT-EMG-A corpus considering 20 trials (N = 20).

K	Mean	σ	Min	Max
1	47.73	6.34	32.95	59.09
2	40.50	6.90	27.50	51.25
3	34.24	5.56	27.78	48.61
4	30.70	6.21	20.31	40.63
5	26.61	4.33	16.07	35.71
6	26.67	6.33	18.75	39.58
7	25.25	6.43	15.00	35.00
8	**22.50**	7.42	**9.38**	37.50
9	25.62	6.38	12.50	37.50

Based on Table 4, we find the best result for K = 8 having an average WER of 22.50%. The best run was achieved for K = 8 with a 9.38% WER.

The classification results for the PT-EMG-B corpus are described in Table 5. For this corpus we find the best result for K = 10 with an average WER of 42.29% and the best run for K=11 with a WER of 31.25%.

Table 5. Surface EMG WER classification results for the PT-EMG-B corpus considering 20 trials (N = 20).

K	Mean	σ	Best	Worst
1	64.10	5.55	50.89	75.00
2	56.87	5.97	44.23	65.38
3	53.38	6.09	43.75	63.54
4	51.19	5.10	43.18	61.36
5	50.50	6.89	36.25	66.25
6	50.06	6.50	40.27	61.11
7	47.89	4.33	39.06	53.12
8	47.14	6.61	35.71	57.14
9	45.72	5.42	33.33	54.16
10	**42.29**	4.53	35.00	52.50
11	43.13	6.92	**31.25**	56.25
12	42.88	6.66	33.33	54.17

By analysing the WER values across K on both corpus, the verified trend (WER decreases when K increases), indicates that increasing the amount of observations in the training set might be beneficial to the applied technique. Regarding the difference in the results with the two corpora, an absolute difference of almost 20.0% and a relative difference of 46.80% are verified between the best mean results. If we compare the best run results, then an absolute difference of 21.87% and a relative difference of 70.0% are verified. Results indicate a major discrepancy between the results from both corpora, hence an error analysis was also performed. If we examine the error represented by the confusion

matrix of the best run for the PT-EMG-B corpus (depicted in Figure 4), we verify that most of the misclassifications will occur in the nasal pairs. In this case errors can be found between the following pairs: [kɐ̃tu] as [katu] and vice-versa; [pɐ̃tɐ] as [petɐ]; [mɐ̃tu] as [matu]; and [titu] as [tĩtu]. In Table 6, the average error percentage for all minimal pairs in Table 3, is presented for each value of K. These results show that 25.4% of the results, 50.8% of the total error, occur between the analysed minimal pairs. This result, leads us to conclude that the current techniques for silent speech recognition based on sEMG, will present a degraded performance when dealing with languages with nasal characteristics like EP.

Table 6: Error analysis for the several values of K. The "Correct" column contains the percentage of correct classifications, e. g. observation *cato* was classified as *cato*. The "Minimal Pair Error" column represents the percent of observations classified as its pair, e. g. observation *cato* was classified as *canto*. The "Remaining Error" column presents the remaining classification errors.

K	Correct (%)	MP Error (%)	Remaining Error (%)
1	35.89	21.42	42.67
2	43.12	24.18	32.69
3	46.61	24.58	28.80
4	48.80	26.07	25.11
5	49.50	27.31	23.18
6	49.93	27.29	22.77
7	52.10	26.01	21.87
8	52.85	26.69	20.44
9	54.27	26.25	19.47
10	57.12	25.00	17.87
11	56.87	24.53	18.59
Mean	**49.74**	**25.40**	**24.87**

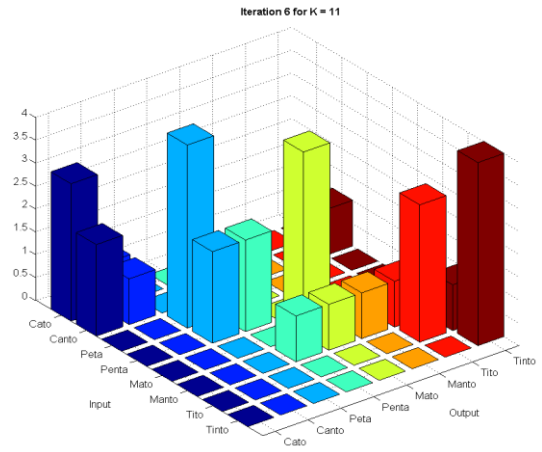

Figure 4: Confusion matrix for the best run (K=11 and N=6).

4.6 Discussion

The results from the PT-EMG-A corpus (average WER of 22.50% and best WER of 9.38%) show a slightly worst accuracy when compared with the latest state-of-the-art results for EMG-based recognition (Schultz and Wand, 2010), considering that a much larger vocabulary was used. This may be explained by the low number of observations as demonstrated by the improvement verified when K increases and by hardware limitations in terms of unipolar configurations and used pairs of sensors. In the results from the PT-EMG-B corpus we verify a relative difference towards the WER results from the first corpus that in the best run case reaches the 70%. Considering that the first four words of the PT-EMG-A corpus were repeated and that the only difference for the remaining words is the presence of nasality in one of the phones, it all points to inadequate handling of the nasality phenomena by sEMGs, as a potential error source. Additionally, the error analysis also indicates that 50.8% of the total error occurs in the minimal pairs, confirming what was stated before.

It´s interesting to mention that the results verified in this experiment in terms of nasality detection, were similar to the ones achieved for Visual Speech Recognition (Freitas et al., 2011) using the same corpora and similar circumstances.

These results suggest that all stages of the SSI recognition process should be reviewed in order to overrun the challenge presented by the nasality phenomena verified in European Portuguese. It needs to be analysed if the muscles involved in the nasal process can be detected by the surface EMG sensors and if the features that characterize the signal correctly, represent this process well. In terms of classification, increasing the number of observations would also allow us to use a more robust classification method such as, Hidden Markov models.

5 CONCLUSIONS

In this paper we have analysed the adoption of existent state-of-the-art techniques in terms of silent speech recognition for the case of a Western Romance language: European Portuguese. We began by describing the current research status in this topic, exposing the present-day challenges and the relation of these challenges with the adoption to a new Romance language. We have also presented the most relevant characteristics of the considered

language, giving emphasis to nasality, one of the present challenges in SSIs. In view of these facts, an experiment towards silent speech recognition in European Portuguese was built. The experience focused on the identification of nasality, and two separate corpora were used. The first corpus, composed of 8 European Portuguese words, allowed us to validate the used framework. The verified results were similar to state-of-the-art (Schultz and Wand, 2010) (best WER 9.38%), if we consider the low number of observations used and some hardware restrictions. The second corpus, was composed by 4 minimal pairs, where the presence or absence of nasality in one of its phones was the only difference. Results from the second corpus show a relative difference towards the results from the first that, in the best run case, reaches 70%. Error analysis from these results indicates that 50.8% of the total error is verified in the nasal pair. These results allow us to conclude that when considering a language with strong nasal characteristics, such as European Portuguese, for a SSI based on sEMG, performance degradation will be verified due to difficulties of the technique in distinguishing nasal phones from oral ones, motivating further research on this topic.

5.1 Future work

Regarding future work we can identify several main research paths to explore. One of them is, of course, nasality. Although the behaviour of the articulators during nasals is not yet clear for EP, we believe that its detection is not impossible. For this reason, we intent to analyse if it is possible to identify nasality in sEMG signals, by sensing specially selected set of muscles, with a stronger association to nasal sounds, or if we can extract nasality information from the current signals, considering other features beyond the time-domain analysis. Another path to consider is the use of parallel modalities appropriate for SSI, which could improve the detection of nasality in European Portuguese, such as Visual Speech Recognition and Ultrasonic Doppler sensing. Particularly, the non-invasive characteristics and the promising results seen in previous works of the Ultrasonic Doppler sensing (Srinivasan et al., 2010) modality open the possibility of a multimodal SSI that addresses the nasality detection problem. A solution for this issue would enable language expansion for this type of interfaces.

ACKNOWLEDGEMENTS

This work was co-funded by the European Union under QREN 7900 LUL - Living Usability lab (http://www.livinglab.pt/) and Golem (ref.251415, FP7-PEOPLE-2009-IAPP).

REFERENCES

Betts, B.J. Binsted, K. Jorgensen, C., 2006. Small-vocabulary speech recognition using surface electromyography. *Journal Interacting with Computers*, vol. 18, pp. 1242–1259.

Chan, A.D.C. Englehart, K. Hudgins, B. and Lovely, D.F., 2001. Hidden Markov model classification of myoelectric signals in speech. *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, pp. 1727–1730.

Chan, A.D.C., 2003. Multi-expert automatic speech recognition system using myoelectric signals. *Ph.D. Dissertation*, Department of Electrical and Computer Engineering, University of New Brunswick, Canada.

De Luca, C.J., 1979. Physiology and mathematics of myoelectric signals. *IEEE Transactions on Biomedical Engineering*, vol. BME-26, no. 6, pp. 313–325.

Denby, B. Schultz, T. Honda, K., Hueber, T. Gilbert, J.M. and Brumberg, J.S., 2010. Silent speech interfaces. *Speech Communication*, v.52 n.4, April 2010, pp. 270-287.

Dias, M. S. Bastos, R. Fernandes, J. Tavares, J. and Santos, P., 2009. Using Hand Gesture and Speech in a Multimodal Augmented Reality Environment, *GW2007*, LNAI 5085, pp.175-180.

Freitas, J. Teixeira, A. Dias M. S. and Bastos, C., 2011. Towards a Multimodal Silent Speech Interface for European Portuguese, *Speech Technologies*, Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech.

Gerdle, B. Karlsson, S. Day, S. Djupsjöbacka, M., 1999. Acquisition, processing and analysis of the surface electromyogram. in *Modern Techniques in Neuroscience*, U. Windhorst and H. Johansson, Eds. Berlin: Springer Verlag, pp. 705–755.

Hardcastle, W. J., 1976. Physiology of Speech Production - An Introduction for Speech Scientists. *Academic Press*, London.

Herff, C. Janke, M. Wand, M. Schultz, T., 2011. Impact of Different Feedback Mechanisms in EMG-based Speech Recognition. *Interspeech 2011*. Florence, Italy.

Jorgensen, C. Lee, D. and Agabon, S., 2003. Sub auditory speech recognition based on EMG signals. In *Proc. Internat. Joint Conf. on Neural Networks* (IJCNN), pp. 3128–3133.

Jorgensen, C. Binsted, K., 2005. Web browser control using EMG based sub vocal speech recognition. In: *Proc. 38th Annual Hawaii Internat. Conf. on System Sciences. IEEE*, pp. 294c.1–294c.8.

Jorgensen, C. and Dusan, S., 2010. Speech interfaces based upon surface electromyography, *Speech Communication*, Volume 52, Issue 4, pp. 354-366.

Jou, S. Schultz, T. Walliczek, M. Kraft, F. and Waibel, A., 2006. Towards Continuous Speech Recognition Using Surface Electromyography. I*nternational Conference of Spoken Language Processing, Interspeech 2006* - ICSLP, Pittsburgh, PA.

Jou, S. Schultz, T. Waibel, A., 2007. Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007*, Honolulu, Hawaii, US.

Junqua, J.-C. Fincke, S. and Field, K., 1999. The Lombard effect: a reflex to better communicate with others in noise. In *Proc. IEEE Internat Conf. on Acoust. Speech Signal Process.* (ICASSP). pp. 2083–2086.

Maier-Hein, L. Metze, F. Schultz, T. and Waibel, A., 2005. Session independent non-audible speech recognition using surface electromyography, *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, pp. 331–336.

Magen, H.S, 1997. The extent of vowel-to-vowel coarticulation. In *English, J. Phonetics* 25 (2), pp. 187–205.

Manabe, H. Hiraiwa, A. Sugimura, T., 2003. Unvoiced speech recognition using EMG-mime speech recognition. In: *Proc. CHI, Human Factors in Computing Systems*, Ft. Lauderdale, Florida, pp. 794–795.

Manabe, H. Zhang, Z., 2004. Multi-stream HMMfor EMG-based speech recognition. In: *Proc. 26th Annual International Conf. of the IEEE Engineering in Medicine and Biology Society*, 1–5 September 2004, San Francisco, California, Vol. 2, pp. 4389–4392.

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, December 2008, pp. 925–952.

Pêra, V. Moura, A. and Freitas, D., 2004. LPFAV2: a new multi-modal database for developing speech recognition systems for an assistive technology application, In *SPECOM-2004*, pp. 73-76.

Plux Wireless Biosignals, 2011. Portugal, [online] Available at: http://www.plux.info/ [Accessed 8 September 2011].

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.

Sá, F. Afonso, P. Ferreira, R. and Pera, V., 2003. Reconhecimento Automático de Fala Contínua em Português Europeu Recorrendo a Streams Audio-Visuais. In *The Proceedings of COOPMEDIA'2003 - Workshop de Sistemas de Informação Multimédia, Cooperativos e Distribuídos*, Porto, Portugal.

Schultz, T. and Wand. M., 2010. Modeling coarticulation in large vocabulary EMG-based speech recognition. *Speech Communication*, Vol. 52, Issue 4, April 2010, pp. 341-353.

Seikel, J. A. King, D. W. and Drumright, D. G., 2010. *Anatomy and Physiology for Speech, Language, and Hearing*, 4rd Ed., Delmar Learning.

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

Strevens, P., 1954. Some observations on the phonetics and pronunciation of modern Portuguese, Rev. Laboratório Fonética Experimental, Coimbra II, pp. 5–29.

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

Teixeira, A. and Vaz, F., 2000. Síntese Articulatória dos Sons Nasais do Português. *Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada* (PROPOR), ICMC-USP, Atibaia, São Paulo, Brasil, 2000, pp. 183-193.

Teixeira, A. Moutinho, L. C. and Coimbra, R.L., 2003. Production, acoustic and perceptual studies on European Portuguese nasal vowels height. In *Internat. Congress Phonetic Sciences* (ICPhS), pp. 3033–3036.

Toda, T. Nakamura, K. Nagai, T. Kaino, T. Nakajima, Y. and Shikano, K., 2009. Technologies for Processing Body-Conducted Speech Detected with Non-Audible Murmur Microphone. In P*roceedings of Interspeech 2009*, Brighton, UK.

Tran, V.-A Bailly, G. Loevenbruck, H. and Toda, T., 2009. Multimodal HMM-based NAM to-speech conversion. In *Proceedings of Interspeech 2009*, Brighton, UK.

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.

Wand, M. and Schultz, T., 2009. Towards Speaker-Adaptive Speech Recognition Based on Surface Electromyography. *In Proc. Biosignals*, pp. 155-162, Porto, Portugal.

Wand, M. Schultz, T., 2011a. Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition, *Interspeech 2011*, Florence, Italy.

Wand, M. Schultz, T., 2011b. Analysis of Phone Confusion in EMG-based Speech Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011*, Prague, Czech Republic.

Wand, M. Schultz, T., 2011c. Session-Independent EMG-based Speech Recognition. *International Conference on Bio-inspired Systems and Signal Processing 2011, Biosignals 2011*, Rome, Italy.

Wilpon, J. G. and Jacobsen, C. N., 1996. A Study of Speech Recognition for Children and the Elderly. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Atlanta, p. 349.