We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us? Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

6

Towards a Multimodal Silent Speech Interface for European Portuguese

João Freitas[1,2,3], António Teixeira[3], Miguel Sales Dias[1,2] and Carlos Bastos[3]
[1]*Microsoft Language Development Center, Tagus Park, Porto Salvo,*
[2]*ISCTE-Lisbon University Institute/ADETTI-IUL,*
[3]*Departamento de Electrónica Telecomunicações e Informática/IEETA,*
Universidade de Aveiro
Portugal

1. Introduction

Automatic Speech Recognition (ASR) in the presence of environmental noise is still a hard problem to tackle in speech science (Ng et al., 2000). Another problem well described in the literature is the one concerned with elderly speech production. Studies (Helfrich, 1979) have shown evidence of a slower speech rate, more breaks, more speech errors and a humbled volume of speech, when comparing elderly with teenagers or adults speech, on an acoustic level. This fact makes elderly speech hard to recognize, using currently available stochastic based ASR technology. To tackle these two problems in the context of ASR for Human-Computer Interaction, a novel Silent Speech Interface (SSI) in European Portuguese (EP) is envisioned.

A SSI performs ASR in the absence of an intelligible acoustic signal and can be used as a human-computer interface (HCI) modality in high-background-noise environments such as, living rooms, or in aiding speech-impaired individuals such as elderly persons (Denby et al., 2010). By acquiring sensor data from elements of the human speech production process – from glottal and articulators activity, their neural pathways or the brain 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.

The aim of this work, is to identify possible multimodal solutions that address the issues raised by adapting existing work on SSIs to a new language. For that motive, a brief state-of-the-art review including current SSI techniques, along with the EP language characteristics that have implications in the capabilities of the different techniques, is provided. The work here presented, describes an initial effort in developing a SSI HCI modality which targets all users (universal HCI) including elderly, for EP. This SSI will then be used as a HCI modality to interact with computing systems and smartphones, respectively, in indoor home scenarios and mobility environments. This paper focuses on the approaches of Visual Speech Recognition (VSR) and Acoustic Doppler Sensors (ADS) for speech recognition, evaluating novel methodologies in order to cope with EP language characteristics. The work here presented describes the initial stage of a novel approach to VSR based on feature tracking using the FIRST algorithm for information extraction. FIRST stands for Feature

Invariant to Rotation and Scale Transform and has been successfully applied to other areas of research (Bastos & Dias, 2008). Regarding the Doppler approach, several examples of this technique are presented for the first time for EP, analyzing important language characteristics such as, nasality.

This paper is structured as follows: next section presents a brief critical review of the state-of-the-art, updating and complementing in depth recent reviews regarding SSIs (Denby et al., 2010); sections 3 and 4 provide a state-of-the art background for the two selected approaches – VSR and ADS ; in section 5 several alternative approaches are described; section 6 discusses the specificities of EP and how these relate to SSI; section 7 presents two experiments aiming at assessing the potential/limitations of some of the techniques in EP specific problems, particularly on identifying the nasal sounds (vowels and consonants).

2. Silent speech interfaces

A Silent Speech Interface comprises a system that interprets human signals other than the audible acoustic signal enabling speech communication (Denby et al., 2010). The SSI paradigm, in the context of speech technologies related research, has rose from the need of aiding speech-handicapped, communications that require the absence of sound or noise such as the ones used in military scenarios, or to increase the robustness of speech recognition in environments with a high level background noise. A SSI system is commonly characterized by the acquisition of information from the human speech production process such as, articulators, facial muscle movement or brain activity. Although further research is still needed in this field, the concept present in a SSI holds a potential solution for a more natural interface for those with disabilities at the vocal tract level, such as people who have undergone a laryngectomy and seniors with speaking difficulties. Informally, one can say that a SSI extends the human speech production model by the use of sensors, ultrasonic waves or vision. 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).

Outside biomedical research, the communications sector, has known great expansion in the last years, and has also become interested on SSIs. The increasing number of mobile devices worldwide has spawned the need for privacy in cell phone conversations. In public environments where silence is often required such as, meetings, cinema or talks, someone talking on the cell phone is usually considered annoying, thus providing the ability to perform an urgent call in this situations has become a point of common interest. Likewise, disclosure of private conversations can occur by performing a phone call in public places, which leads to embarrassing situations from the caller point of view or even to information leaks.

2.1 SSIs and the speech production chain

The speech production model can divided into several stages. According to Levelt (Levelt, 1989), the communicative intention phase is the first phase of each speech act and consists in converting patterns of goals into messages followed by the grammatical encoding of the preverbal message to surface structure. The next phase of the speech production is the passage from the surface structure to the phonetic plan, which, informally speaking is the sequence of phones that are fed to the articulators. This can be divided between the

electrical impulse fed into the articulators and the actual process of articulating. The final phase consists on the consequent effects of the previous phases.

The existent experimental SSI systems described in the literature, cover extraction of information from all the stages of speech production, from intention to articulation to effects of articulation, as depicted on Fig. 1. The current approaches can be divided as follows:

- **Intention level** (brain / Central Nerve System): Interpretation of signals from implants in the speech-motor cortex (Brumberg et al., 2010) , Interpretation of signals from electro-encephalographic (EEG) sensors (Porbadnigk et al., 2009);
- **Articulation control** (muscles): surface Electromyography (sEMG) of the articulator muscles or the larynx (Wand et al., 2009 & Maier-Hein et al., 2005);
- **Articulation** (articulators): Capture of the movement of fixed points on the articulators using Electromagnetic Articulography (EMA) sensors (Fagan et al., 2008); Real-time characterization of the vocal tract using ultra-sound (US) and optical imaging of the tongue and lips (Denby & Stone, 2004; Hueber et al., 2008); Capture movements of a talker's face through ultrasonic sensing devices (Srinivasan et al., 2010; Kalgaonkar et al. 2008);
- **Articulation effects**: Digital transformation of signals from a Non-Audible Murmur (NAM) microphone (a type of stethoscopic microphone) (Toda et al., 2009) , Analysis of glottal activity using electromagnetic (Ng et al., 2000; Quatieri et al., 2006), or vibration (Patil et al., 2010) sensors;

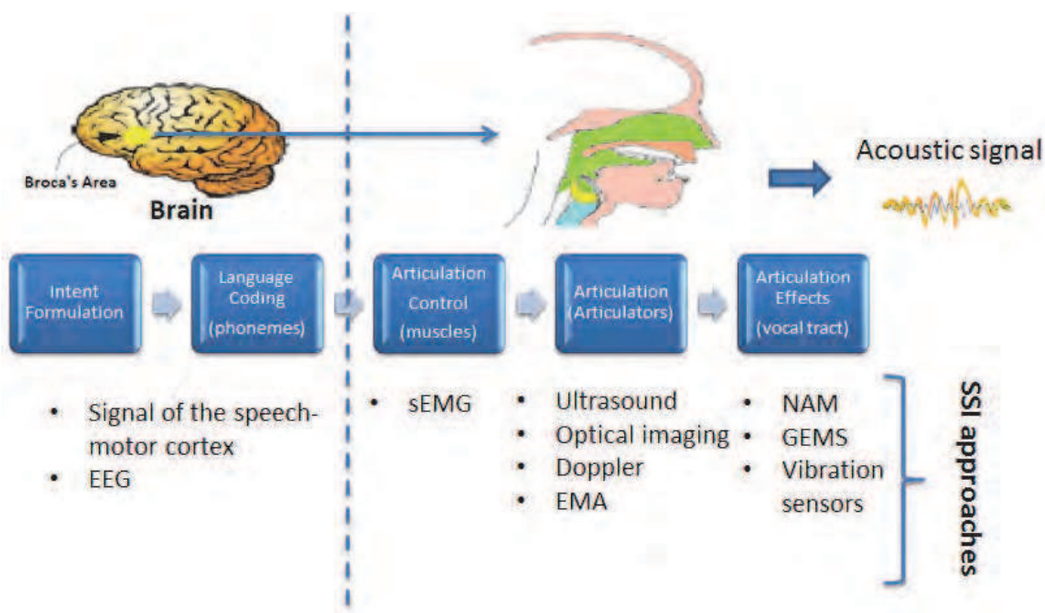


Fig. 1. Phased speech production model with the correspondent SSIs

The taxonomy presented above and illustrated in Fig. 1 allows associating each type of SSI to a stage of the human speech production model providing a better understanding from where the speech information is extracted.

Existent SSI approaches only consider VSR used as a complement for other approaches such as, ultrasound imaging. Furthermore, only lips are considered in the region of interest (ROI), not taking into account information extracted from jaws and cheeks.

2.2 Challenges

In the work presented by Denby and coworkers (Denby et al., 2009) several challenges to the development of SSIs can be found, such as:

- Sensor positioning – In order to achieve speaker independence and higher usability rates a more robust method for positioning the sensors must be found. Currently, the results of several approaches such as, EMA, EMG and EEG, are highly sensible to the position of the sensors requiring previous training or very accurate deployment.
- Speaker Independence – Many of the SSIs extracted features depend on the user's anatomy, experience and even his/her synaptic coding.
- Lombard and silent speech effects – The effects adverting from silent speech articulation and the Lombard effect (different articulation when no auditory feedback is provided) are not yet clear and require further investigation.
- Prosody and nasality – The extraction of speech cues for prosody (in systems where output synthesis is envisaged) and nasality in SSIs is also an issue. Due to the modified or absent speech signal, the information for these parameters must be obtained by other means.

These categories represent areas in this field where further research is required in order to reach an optimal SSI solution. In the work here presented we focus on a new challenge of adapting an SSI to a new language – European Portuguese – which involves addressing some of the issues referred above, such as, nasality.

3. Automatic visual speech recognition

Automatic Speech Recognition (ASR) has suffered a considerable evolution in the last years, especially, in well-defined applications with controlled environments. However, results clearly show that speech recognizers are still inferior to humans and that the human speech perception clearly outperforms state-of-the art ASR systems (Lippman, 1997). Likewise, in order to achieve a pervasive and natural human-computer interface further improvements are needed in channel robustness, ASR in uncontrolled situations and speakers (e.g. elderly) with environment noise (Potamianos et al., 2003).The human speech perception is bimodal in nature, and the influence of the visual modality over speech intelligibility has been demonstrated by the McGurk effect (Stork & Hennecke, 1996; MacGurk & MacDonald, 1976), which states the following: Vision affects the performance of the human speech perception because it permits to identify the source location; it allows a better segmentation of the audio signal; and it provides information about the place of articulation, facial muscle and jaw movement (Potamianos et al., 2003). This fact has motivated the development of audio-visual ASR (AV-ASR) systems and automatic visual speech recognition (AVSR) systems, as depicted on Fig. 2.

Fig. 2. Visual Speech Recognition system pipeline.

In visual-only ASR systems, a video composed of successive frames is used as an input for the system. Relatively to audio front-end, the VSR system adds a new step before the feature extraction, which consists of segmenting the video and detect the location of the speaker's face, including the lips. After this estimation, suitable features can be extracted. The majority of the systems that use multiple simultaneous input channels such as, audio plus video, have a better performance when compared with systems that depend on a single visual or audio only channel (Liang et al., 2010). This has revealed to be true for several languages such as, English, French, German, Japanese and Portuguese; and for various cases such as, nonsense words, isolated words, connected digits, letters, continuous speech, degradation due to speech impairment, etc. (Potamianos et al., 2003).

3.1 Feature extraction techniques

According to the literature (Liang et al., 2010) there are three basic methodologies to extract features in a face and lip reading system: appearance-based; shape-based; or a fusion of both. The first method is based on information extracted from the pixels in the whole image or from some regions of interest. This method assumes that all pixels contain information about the spoken utterance, leading to high dimensionality issues. Shape-based approaches base the extraction of features in the lip's contours and also parts of the face such as, cheek and jaw. This method uses geometrical and topological aspects of the face in order to extract features, like the height, width and area of the mouth image moment descriptors of the lip contours, active shape models or lip-tracking models. Shape-based methods require accurate and reliable facial and lip feature detection and tracking, which reveal to be complex in practice and hard at low image resolution (Zhao et al., 2009). The third method is a hybrid version of the first and second methods and combines features from both previous methodologies either as a joint shape appearance vector or as a cooperative statistical model learned from both sets of features. Appearance-based methods, due to its simplicity and efficiency, are the most popular (Liang et al., 2010). A comprehensive overview of these methodologies can be found in (Potamianos et al., 2003).

The challenge in extracting features from video, resides in collecting required information from the vast amounts of data present in image sequences. Each video frame contains a large number amount of pixels and is obviously too large to model as a feature vector. In order to reduce dimensionality and to allow better feature classification, techniques based on linear transformations are commonly used such as, PCA - Principal Component Analysis, LDA- Linear Discriminant Analysis, DCT- Discrete Cosine Transform and DWT - Dynamic Time Warping, Haar transforms, LSDA- Locality Sensitive Discriminant Analysis, or a combination of these methods (Potamianos et al., 2003; Liang et al., 2010). In our work we have followed an appearance-based approach using vision-based feature extraction and tracking. This type of approach has proliferated in the area of computer vision due to its intrinsic low computational cost, allowing real time solutions. To fully address our problem, we need a robust feature extraction and tracking mechanism and the Computer Vision community provides us different alternatives, such as (Harris & Stephens, 1988), SIFT – Scale Invariant Feature Transform (Lowe, 2004), PCA-SIFT (Ke & Sukthankar, 2004), SURF - Speeded Up Robust Features (Bay et al., 2006) or FIRST-Fast Invariant to Rotation and Scale Transform (Bastos & Dias, 2009). In terms of VSR, the concepts behind these techniques have been used for the elimination of dependencies on affine transformations by (Gurbuz et al., 2001) and promising results, in terms of robustness, have been achieved. These methods

have shown high matching accuracy on the presence of affine transformations. However, some limitations in real-time applications were found. For example, an objective analysis in (Bastos & Dias, 2009), showed that SURF took 500 milliseconds to compute and extract 1500 image features on images with resolution of 640x480 pixel, while PCA-SIFT took 1300 ms, SIFT took 2400 ms and FIRST only 250 ms (half of the second most efficient, SURF). As for matching the same 1500 features against themselves, the figures for SURF, PCA-SIFT, SIFT and FIRST were, respectively, 250 ms, 200 ms, 800 ms and 110 ms, as observed by (Bastos & Dias, 2009).Given these results, we have selected FIRST as the technique to be used in this work to extract and match features, since the real-time requirement is essential in practical VSR systems.

The selected approach, FIRST, can be classified as a corner-based feature detector transform for real-time applications. Corner features are based on points and can be derived by finding rapid changes in edge's direction and analyzing high levels of curvature in the image gradient. FIRST features are extracted using minimum eigenvalues and are made scale, rotation and luminance invariant (up to some extent of course, since with no luminance, no vision in the visible domain is possible), using real-time computer vision techniques. A more detailed description of the algorithm is provided in section 7.1. The FIRST feature transform was mainly developed to be used in application areas of Augmented Reality, Gesture Recognition and Image Stitching. However, to our knowledge this approach has not yet been explored in VSR and the potential for extracting information, such as nasality, using this technique is still novelty.

4. Doppler signals for SSI

Ultrasound Doppler sensing of speech is one of the approaches reported in the literature that is also suitable for implementing a SSI (Srinivasan et al., 2010). This technique is based on the emission of a pure tone in the ultrasound range towards the speaker's face that is received by an ultrasound sensor tuned to the transmitted frequency. The reflected signal will contain Doppler frequency shifts proportional to the movements of the speaker's face. Based on the analysis of the Doppler signal, patterns of movements of the facial muscles, lips, tongue, jaw, etc., can be extracted (Toth et al., 2010). The ADS have been previously used in voice activity detection (Kalgaonkar et al., 2007), speaker identification (Kalgaonkar & Raj, 2008), speech recognition (Zhu et al., 2007; Srinivasan et al., 2010) and synthesis (Toth et al., 2010). When using ADS for speech recognition, the sensor has been placed either at 6-8 inches or 16 inches from the speaker, being the second more realistic. The results for ultrasound-only approaches are still far from audio-only performance. Best results in speech recognition (Srinivasan et al., 2010) show a recognition accuracy of 33% in speaker independent digit recognition, using HMMs trained with recordings of 6 speakers, revealing viability and margin for improvement of this approach.

4.1 The Doppler Effect

The Doppler Effect is the change in frequency of an emitted wave perceived by a listener moving relative to the source of the wave. If we consider the scenario depicted on Fig. 3, where the source T emits a wave with frequency f_0 that is reflected by the moving object, in this case the speaker's face. The reflected signal is then given by Eq. 1, with v being the velocity of the moving object based on the transmitter T and v_s is the velocity of the sound in the medium.

$$f = f_0\left(\frac{v_s + v}{v_s - v}\right) \tag{1}$$

Since articulators move at different velocities when a person speaks, the reflected signal will have multiple frequencies each one associated with the moving component (Toth et al., 2010).

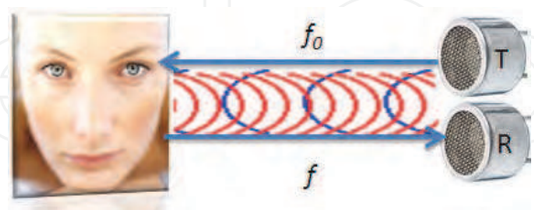


Fig. 3. Doppler Effect representation (T – Transmitter, R- Receptor).

To our knowledge, only recently Doppler has been applied to speech recognition in (Srinivasan et al., 2010) and no research of this technique regarding EP has yet been published. Below, a first example of the Doppler signal and the correspondent audio signal applied to an EP word *canto* [kɐ̃tu] (corner), is depicted on Fig. 5.

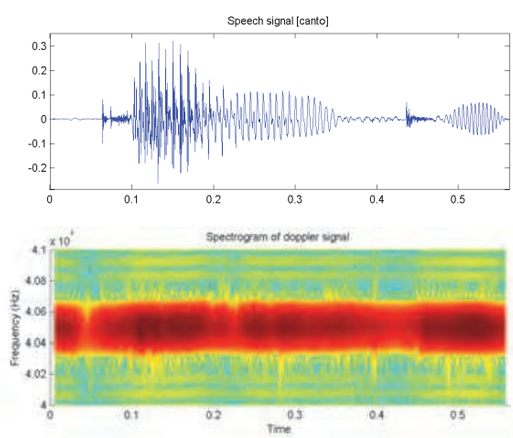


Fig. 5. Audio signal (above) and spectrogram of the Doppler signal (below) for the word *canto*.

This approach, like it was stressed before, still has margin for improvement especially when applied for Silent Speech recognition and the potential for detecting characteristics such as, nasality is still unknown. For that reason, and for being an example of very recent work, a first approach to EP using this technology is described on section 7.2. Future research in this area will need to address issues such as, changes in pose and distance of the speaker variation that affect the ultrasound performance.

5. Other SSI approaches

In this section alternative technologies to ADS and VSR, for silent speech recognition, are described according to the several stages of the speech production model and it is explained in what way information can be collected at all stages, starting by the intention phase to the articulation effects.

5.1 Brain computer interfaces used for silent speech recognition

The goal of a Brain Computer Interface (BCI) is to interpret thoughts or intentions and convert them into a control signal. The evolution of cognitive neuroscience, brain-imaging and sensing technologies has provided means to understand and interpret the physical processes in the brain. BCI's has a wide scope of application and can be applied to several problems like, assistance to subjects with physical disabilities (e.g. mobility impairments), detection of epileptic attacks, strokes, or to control computer games (Nijholt et al., 2008). A BCI can be based in several types of changes that occur during mental activity, such as, electrical potentials, magnetic fields, or metabolic/hemodynamic recordings. Current SSI approaches have been based on electrical potentials, more exactly on the sum of the postsynaptic potentials in the cortex. Two types of BCI's have been used for unspoken speech recognition, one invasive approach based on the interpretation of signals from intra-cortical microelectrodes in the speech-motor cortex and a non-invasive approach based on the interpretation of signals from electro-encephalographic sensors. Unspoken speech recognition tasks have also been tried based on magnetoencephalograms (MEG), which measures the magnetic fields caused by current flows in the cortex. However, results have shown no significant advantages over EEG-based systems (Suppes et al., 1997). A wide overview of BCI's can be found in (Nijholt et al., 2008).

5.1.1 Intra-cortical microelectrodes

Due to the invasive nature, increased risk and medical expertise required, this approach is only applied as a solution to restore speech communication in extreme cases such as, subjects with the locked-in syndrome medically stable and presenting normal cognition. When comparing with EEG sensors this type of systems present a better performance enabling real-time fluent speech communication (Brumberg et al., 2010). This technique consists in the implantation of an extracellular recording electrode and electrical hardware for amplification and transmission of brain activity. Relevant aspects of this procedure include: the location for implanting the electrodes; the type of electrodes; and the decoding modality. Results for this approach in the context of a neural speech prosthesis show that a subject is able to correctly perform a vowel production task with an accuracy rate up to 89% after a several month training period (Brumberg et al. 2009).

5.1.2 EEG sensors

A SSI based on unspoken speech is particularly suited for subjects with physical disabilities such as, the locked-in syndrome. The term unspoken speech refers to the process where the subject imagines speaking a given word without moving any articulatory muscle or producing any sound. Results from this approach have achieved accuracies significantly above chance (4 to 5 times higher) and indicate that the Broca's and Wernicke's areas as the most relevant in terms of sensed information (Wester & Schultz, 2006). Other studies (DaSalla et al., 2009) were performed using vowel speech imagery, regarding the classification of the vowels /a/ and /u/ and achieved overall classification accuracies ranged from 68 to 78%, indicating the use of vowel speech as a potential speech prosthesis controller.

5.2 Surface EMG

According to the speech production model, the articulator's muscles are activated through small electrical currents in the form of ion flows, originated in the central and peripheral nervous systems. The electrical potential generated by the resistance of muscle fibers, during

speech, leads to patterns that occur in the region of the face and neck, which can be measured by a bioelectric technique called surface electromyography. This technique consists in the study of the muscle activity through its electrical properties. Currently, there are two sensing techniques to measure electromyography signals: invasive indwelling sensing and non-invasive sensing. In this work, as in most of the sEMG-based ASR research the latter approach is used (Betts & Jorgensen, 2006). Since this approach only relies on the analysis of the resulting myoelectric signal pattern related with muscle activity, it allows overcoming the main limitations of current ASR technology such as, robust ASR in the presence of environmental noise, and use of ASR near bystanders and for private actions (Denby et al., 2010). This technology has 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 & Jorgensen, 2006). Latest studies show that recognition rates up to 90% can be achieved, for a 100-word vocabulary in a speaker dependent scenario (Jou et al. ,2007; Schultz & Wand, 2010), or for small vocabularies scenarios where only one pair of surface electrodes is used (Jorgensen et al., 2003).

5.3 Electromagnetic articulography

By monitoring the movement of fixed points in the articulators using EMA sensors, this approach collects information from the articulation stage (referred on Fig. 1) of the speech production model. It is based on implanted coils or magnets attached to the vocal apparatus (Fagan et al., 2008), which can be electrically connected to external equipment or coupled with magnetic sensors positioned around the user's head. The movements of the sensors are then tracked and associated with the correspondent sound. For example, in Fagan (Fagan et al., 2008), magnets were placed on the lips, teeth and tongue of a subject and were tracked by 6 dual axis magnetic sensors incorporated into a pair of glasses. Results from this laboratory experiment show an accuracy of 94% for phonemes and 97% accuracy for words, considering very limited vocabularies (9 words and 13 phonemes).

5.4 Ultrasound and optical imaging of the tongue and lips

One of the limitations found in VSR process described earlier is the visualization of the tongue, a vital articulator for speech production. In this approach this limitation is overcome by placing beneath the chin, an ultrasound transducer, thus providing a partial view of the tongue surface in the mid-sagittal plane (Denby et al., 2010). This type of approaches is commonly combined with frontal and side optical imaging of the user's lips. For this type of systems the ultrasound probe and the video camera are usually fixed to a table or a helmet to ensure that no head movement is performed or that the ultrasound probe is correctly oriented in regard to the palate and the camera is kept at a fixed distance (Florescu et al., 2010). Latest work in the US/Video approach relies on a global coding approach in which images are projected onto a more fit space regarding the vocal tract configuration – the EigenTongues. This technique encodes not only tongue information but also information about other structures that appear in the image such as, hyoid bone and muscles (Hueber et al., 2007). Results for this technique show that for an hour of continuous speech, 60% of the phones are correctly identified in a sequence of tongue and lip images, showing that better performance can be obtained using more limited vocabularies or using isolated word in silent speech recognition tasks, still considering realistic situations (Hueber et al., 2009).

5.5 Non-audible murmur microphones

Non-audible murmur is the term given by research community to low amplitude speech sounds produced by laryngeal airflow noise resonating in the vocal tract (Denby et al., 2010). This type of speech is not perceptible to nearby listeners, but can be detected using the NAM microphone, introduced by (Nakajima et al., 2003). This microphone can be used in the presence of environmental noise, enables some degree of privacy and can be a solution for subjects with speaking difficulties or laryngeal disorders such as, the elderly. The device consists on a condenser microphone covered with soft silicone or urethane elastomer, which helps to reduce the noise caused by friction to skin tissue or clothing (Otani et al., 2008). The microphone diaphragm is exposed and the skin is in direct contact with the soft silicone. This device has a frequency response bandwidth of about 3 kHz, with peaks at 500-800 Hz and some problems concerning small spectral distortions and tissue vibration have been detected. However, the device remains as an acceptable solution for robust speech recognition (Denby et al., 2010). The best location for this microphone was determined by (Nakajima, 2005) to be on the neck surface, more precisely below the mastoid process on the large neck muscle. This technology has also been tried in a multimodal approach in (Tran et al., 2010) where this approach is combined with a visual input. In terms of recognition accuracy, Herucleous (Herucleous et al., 2003) has reported values around 88% using an iterative adaptation of normal-speech to train HMM's, requiring only a small amount of NAM data.

5.6 Electromagnetic and vibration sensors

The development of this type of sensors was motivated by several military programs in Canada, EUA and European Union to evaluate non-acoustic sensors in acoustically harsh environments such as, interiors of military vehicles and aircrafts. In this case, by non-acoustic it is meant that the sound is propagated through tissue or bone, rather than air (Denby et al., 2010). The aim of these sensors is then to remove noise by correlating the acquired signal with the one obtained from a standard close-talk microphone. These sensors have presented good results in terms of noise attenuation with gains up to 20 db (Dupont & Ris, 2004; Quatieri et al., 2006) and word error rate (WER) significant improvements (Jou et al., 2004). It has also been presented by (Quatieri et al., 2006) that these sensors can be used to measure several aspects of the vocal tract activity such as low-energy, low-frequency and events such as, nasality, which is strong characteristic of EP as described in section 6.1. Based on these facts, the use of these technologies is being considered by Advanced Speech Encoding program of DARPA for non-acoustic communication (Denby et al., 2010). These types of sensors can be divided into two categories, electromagnetic and vibration. Regarding electromagnetic sensors the following types can be found: Electroglottograph (EGG); General Electromagnetic Motion System (GEMS) and Tuned Electromagnetic Resonating Collar (TERC). In terms of vibration microphones the following types can be found: Throat Microphone, Bone microphone, Physiological microphone (PMIC) and In-ear microphone.

6. SSI for European Portuguese

The existing SSI research has been mainly developed by groups from EUA (Hueber et al., 2009), Germany (Calliess & Schultz, 2006), France (Tran et al., 2009) and Japan (Toda et al.,

2009), which have focused their experiments on their respective languages. There is no published work for European Portuguese in the area of SSIs, although there are previous research on related areas, such as the use of EMA (Rossato et al., 2006), Electroglotograph and MRI (Martins et al., 2008) for speech production studies, articulatory synthesis (Teixeira & Vaz, 2000) and multimodal interfaces involving speech (Teixeira et al., 2005; 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.

6.1 European Portuguese characteristics

According to Strevens (Strevens, 1954), when one first hears European Portuguese (EP), the characteristics that distinguishes 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), results show that European Portuguese stops, revealed less resistant to co-articulatory effects than fricatives.

6.1.1 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] (quando), [wẽ] (aguentar), [jɐ̃] (fiando), [wĩ] (ruim) and triphthongs [wɐ̃w] (enxaguam).

Nasal vowels in EP diverge from other languages with nasal vowels, such as French, in its wider variation in the initial segment and stronger nasality at the end (Trigo, 1993; Lacerda & Head, 1966).

Doubts still remain regarding tongue positions and other articulators during nasals production in EP, namely, nasal vowels (Teixeira et al., 2003). Martins (Martins et al., 2008) have detected differences at the pharyngeal cavity level and velum port opening quotient when comparing EP and French nasal vowels articulation.

7. Experiments analysis

In our research we have designed experiments to analyze and explore two SSI approaches – VSR and ADS – applied to EP. The paper addresses an important research question, regarding the capability of two different approaches to distinguish nasal sounds from oral ones. To tackle this objective, we have designed a scenario where we want to recognize/distinguish words possibly differing only by the presence or absence of nasality in one of its phones. In EP, nasality can distinguish consonants (e.g. the bilabial stop consonant [p] becomes [m], with nasality creating minimal pairs such as [katu]/[matu]) and vowels (in minimal pairs such as [titu]/[tĩtu]).

7.1 Experiment I – visual speech recognition FIRST-based

The visual speech experiment here presented aims at demonstrating a hypothesis, according to which the skin deformation of the human face, caused by the pronunciation of words, can

be captured by studying the local time-varying displacement of skin surface features distributed across the different areas of the face, where the deformation occurs. We can abstract these image features as particles, and here we are interested in studying the kinematic motion of such particles and specially, its displacements in time, in relation to a given reference frame. Differently from authors that focus their research solely in the analysis of lip deformation (Zhao et al., 2009), in our approach we are interested in other areas of the face, in addition to the lip area. It´s worth recalling that this experiment was designed having in mind its further applicability in real-time speech recognition. Due to previous evidence of the superiority of FIRST, if we compare it SIFT, PCA-SIFT and SURF (Bastos & Dias, 2009), in applications, like augmented reality, that require real-time behavior while keeping sufficient precision and robust scale, rotation and luminance invariant behavior, we have decided to use the FIRST features in our experiment.
The envisioned experiment is divided into the phases depicted in Fig.6.

Fig. 6. The phases of our Visual Speech Recognition experiment.

To put our experiment into practice, we have specified, developed and tested a prototype VSR system. It receives an input video containing a visually spoken utterance. We ask the speaker to be quiet for a few moments, so that we are able to extract FIRST features from the video sequence. After just some frames, the number of detected features stabilizes and we refer to those as the calibration features and their position is stored. The number of calibration features remains constant for the full set of the pronounced words, which, in our case is 40.
After calibration, we assume that the speaker is pronouncing a word and therefore, in each frame, we need to track the new position of each feature in the image plane. In our current experiment we are just performing, in each frame, FIRST feature extraction and subsequently template matching with the calibration features. Further optimizations towards real-time behavior are possible, by using the tracking approach of (Bastos & Dias, 2009), which uses optical flow and feature matching in smaller image regions. If the template matching normalized cross correlation is higher that a predefined threshold, then we assume that the feature was matched and its new u, v image position is updated. Then, the Euclidian distance between the new updated feature position in the current frame and its position in the previous frame, is computed. For each feature, the resulting output will be a law in time of the displacement (distance) of each, relatively to the calibration position. During the feature matching process several outliers may occur and are later removed in a post-processing phase (Fig.6).
In each frame, we are able to compute the displacement of each of the human face surface features that we are tracking. These feature displacements, will then be used as input feature vectors for a following machine classification stage. By analyzing these feature vectors during the full story of the observed word pronunciation and comparing these analysis with the remaining examples, we can chose the one with the closest distance, consequently being

able to classify that observation as a recognized word. The distance is obtained by applying Dynamic Time Warping (DTW) (Rabiner & Juang, 1993). In the following sections, we provide a detailed description of the process.

7.1.1 Feature extraction

The feature extraction process follows the work of (Bastos & Dias, 2008) and is performed using Minimum Eigen Values (MEV), as described in the (Shi & Tomasi, 1994) detector. The reason for choosing this detector is related with its robustness in the presence of affine transformations. For feature extraction, the image is first converted to gray scale. Then, a block of 3x3 pixels is taken at every image position and the first derivatives in the direction of x (D*x*) and y (D*y*) are computed using the Sobel operators O*x* and O*y* (Eq. 2), for convolution with the 3x3 pixels block. The convolution will result in evaluation of the mentioned first derivatives in direction of *x* and *y*. With the computed derivatives, we can construct matrix C, where the sum is evaluated in all elements of the 3x3 block. The Eigen Values are found by computing Eq. 4, where I is the identity matrix and λ the column vector of Eigen Values.

$$O_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad O_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{2}$$

$$C = \begin{bmatrix} \sum D_x^2 & \sum D_x D_y \\ \sum D_x D_y & \sum D_y^2 \end{bmatrix} \tag{3}$$

$$det(C - \lambda I) = 0 \tag{4}$$

Two solutions (λ_1, λ_2) will result from the equation and the minimum Eigen Value (min(λ_1, λ_2)) will be retained. In order to perform feature identification and determine strong corners, a threshold is applied to the resulting MEV's. Only features that satisfy the threshold value of 1% of the global maximum in the current MEV spectrum are selected. Non-maximum suppression is also performed by evaluating if the candidate corner's MEV is the maximum in a neighborhood of 3x3 pixels. After the features position in the image plane have been found, several computer vision techniques are applied in order to make such features scale, rotation and luminance invariant, while at the same time maintaining the efficiency requirements. The algorithm for this procedure is already described in (Bastos & Dias, 2008), and for this reason we will only refer which techniques were used.

To make the FIRST features scale invariant it is assumed that every feature has its own intrinsic scale factor. Based on the results of the Sobel filters, the edges length can be directly correlated with the zooming distance. By finding the main edge length of the derivatives an intrinsic scale factor can be computed. The scale factor will then enable for the intrinsic feature patch to be normalized and consequently make it scale invariant. Only a surrounding area of 7x7 pixels relatively to the feature center is considered in order to deal with other derivatives that may appear resultant from zooming in/out.

In order to achieve rotation invariance the highest value of the feature's data orientation is determined. Assuming that the feature's data is an *n* x *n* gray scale image patch (g_i) centered

at (c_x, c_y), already scale invariant, the function that find the main orientation angle of the feature g_i is given by:

$$\theta(g_i) = b \max(H(g_i)) \tag{5}$$

Where $H(g_i)$ gives the highest value of orientation of g_i based on an orientation histogram composed by b elements (each element corresponds to $360^o/b$ degrees interval). The max function returns the $H(g_i)$ histogram vector index. After obtaining the result of Eq. 5, a rotation of $\theta(g_i)$ degrees is performed to the g_i gray scale patch.
Luminance Invariance is accomplished by using a template matching technique that uses invariant image gray scale templates (Bastos & Dias, 2009). This technique is based on the image average and standard deviation to obtain a normalized cross correlation value between features. A value above 0.7 (70%) is used as correlation factor.

7.1.2 Feature matching

The FIRST feature transformation here presented is not as distinctive as SIFT or PCA-SIFT, for that reason, a method based on feature clustering is used. The method groups features into clusters through a binary identification value with low computation cost. The identification signature is obtained by evaluating three distinct horizontal and vertical regions of Difference of Gaussians patches (Davidson & Abramowitz, 2006). Difference of Gaussians is a gray scale image enhancement algorithm, which involves the subtraction of one blurred version of an original gray scale image from another, which is a less blurred version of the original. The blurred gray scale images are obtained by convolving the original image with Gaussian kernels with different standard deviations. The signature will be composed of 8 digits and only features that correspond to a specified feature's binary signature are matched, thus reducing the overall matching time. For positive or null regions a value of 1 will be assigned. Negative regions will be assigned with 0. When a FIRST feature patch is processed and created, this evaluation is performed and this feature is inserted in the corresponding cluster using the obtained binary identification. When matching a feature, we also compute the binary identification of the candidate feature, which allow us to only match with potential candidates instead of matching with all the calibration features collected in a previous phase. For each feature, when a matching is found, the displacement (standard Euclidian distance (Eq. 6)) is computed between the updated feature position and the initial (calibrated) position, given by:

$$d(p, q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \tag{6}$$

The result from this process is bi-dimensional pattern as the one depicted on Fig.8 for the words *cato* [katu] (cact) and *canto*. The horizontal axis represents the video frame (i.e. time) and the vertical axis represents the 40 features displacements, observed in the respective examples of word pronunciation. These images provide us with a view of how features vary in time for different words.
The input videos for these results were recorded under approximately the same conditions and a clear difference between these two words can be noticed in the patterns. Each feature will have a different behavior in its displacement across time, as depicted on Fig. 9, which shows two features from different human face regions for the same word.

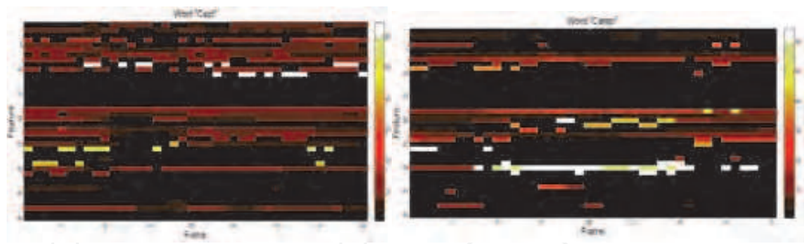


Fig. 8. Image graph for the words *cato* (above) and *canto* (below).

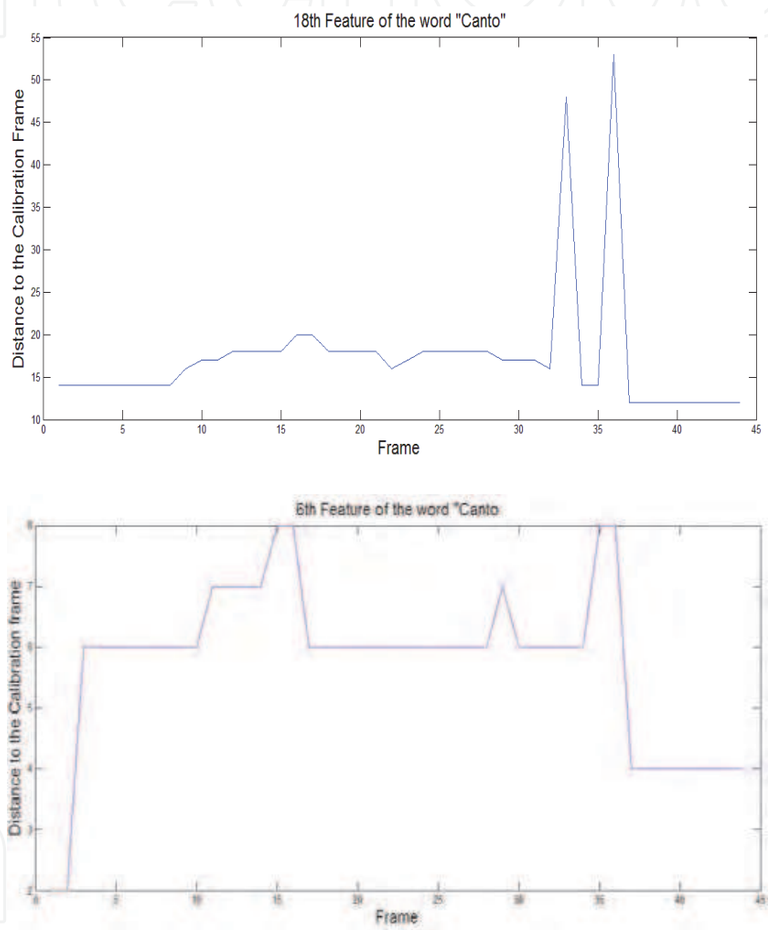


Fig. 9. Temporal behavior of the feature number 18 (right) and number 6 (left) for the word *canto*.

7.1.3 Post-processing

In the post-processing phase outliers have been removed using Chauvenet's criterion (Chauvenet, 1960) with a 0.5 threshold. This approach, although being simple, has shown good results demonstrated by empirical measures. A value is considered to be an outlier when the matching between two features is incorrect, i.e. the matching feature is not the original one. This situation is highlighted on Fig. 10 were the outlier is marked with a red circle on both frames.

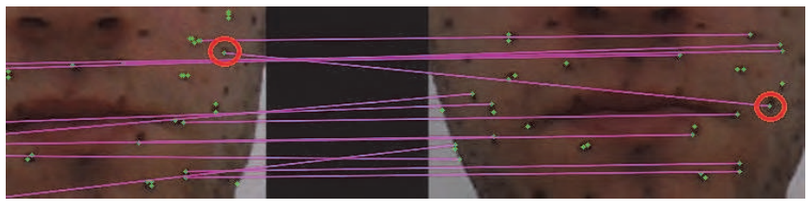


Fig. 10. Matching between two frames with an outlier highlighted by a red circle.

7.1.4 Classification

For this initial stage of research the Dynamic Time Warping (DTW) technique was used to find an optimal match between a sufficient number of observations. DTW 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 Fig. 11 the DTW is applied to several pairs of words observations and we depict the DTW distance results, by means of gray scale coding of such results. For the *cato*/*canto* DTW computation we have, in the horizontal axis, the number of frames of *canto* production, whereas in the vertical axis, we have the number of frames for *cato* pronunciation. These diagrams can be simply interpreted as follows: The similarity between two words is given by the smallest DTW distance between them across time, thus when two words are the same, as shown in the comparison between canto and canto (upper-left graph), the lowest distance will lay in the image's diagonal. The red line represents the lower DTW distances found across time. In the upper-left panel the control case is represented by comparing a word observation with itself originating a straight diagonal line (i.e. all DTW distances in the diagonal are zero). Furthermore, as expected, a certain similarity with the word *Cato* (upper-right panel) can be noticed, since the only difference relies on a nasal phoneme instead of an oral one. It is also visible that the word *tinto* [tĩtu] (red) (bottom-left) presents the highest discrepancy regarding *canto*.

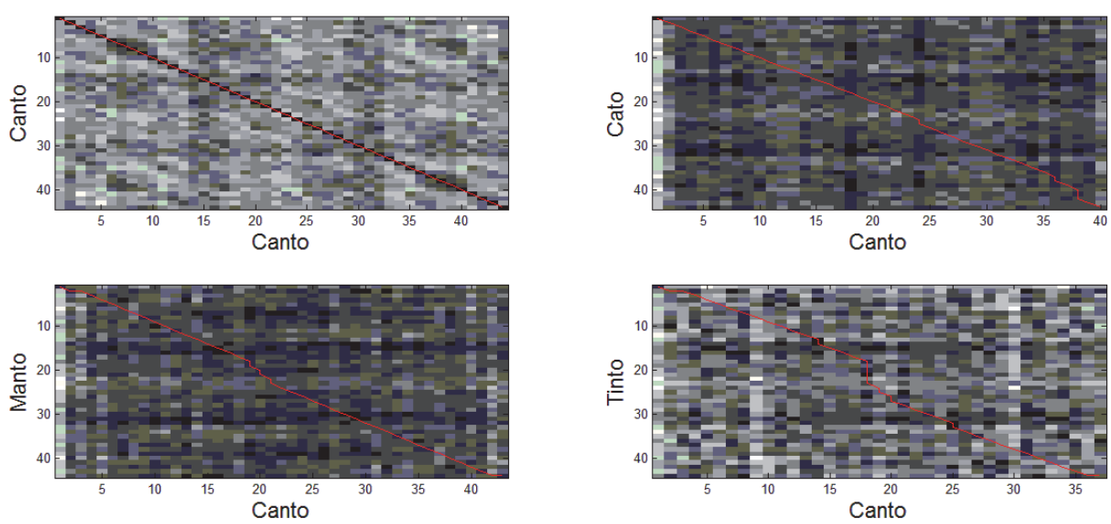


Fig. 11. Distance comparison between the word *canto* and *manto* [mɐ̃tu] (cloak) /*tinto*/*cato* using the Matlab algorithm from (Ellis, 2003).

7.1.5 Experimental setup

The input videos were recorded by one of the authors (JF) using a webcam video of 2 megapixel during daytime with some exposure to daylight. In these videos the user exhibits some spaced facial markings in areas surrounding the lips, chin and cheeks. These fiduciary markers were made with a kohl pencil.

7.1.6 Corpus

For this experiment a database containing 112 video sequences was built from scratch. The videos contain a single speaker uttering 8 different words in European Portuguese, with 14 different observations of each word. These words were not randomly selected and represent four 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*). In Table 1 the pairs of words and their respective phonetic transcription is provided.

Word Pair	Phonetic Transcription
cato/canto	ˈkatu / ˈkɐ̃tu
peta/penta	ˈpetɐ / ˈpẽtɐ
mato/manto	ˈmatu / ˈmɐ̃tu
tito/tinto	ˈtitu / ˈtĩtu

Table 1. Pairs of words used in the VSR experiment

7.1.7 Results and discussion

In order to classify the results the following algorithm based on DTW was applied:

1. Randomly select K observations from each word in the Corpus (see table 1) 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:
   a. Compare each observation with the representative examples.
   b. Select the word that provides the minimum distance.
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.

Considering this algorithm with N = 20 and K varying from 1 to 10, the following results in terms of Word Error Rate (WER) are achieved:

K	Mean	σ	Best	Worst
1	32.98	5.43	25.00	45.19
2	26.93	4.84	19.79	41.67
3	22.95	5.20	15.91	36.36
4	17.94	4.73	11.25	26.25
5	16.04	3.99	9.72	22.22
6	12.81	3.53	7.81	20.31
7	13.04	4.26	3.57	19.64
8	12.08	3.68	6.25	22.92
9	8.63	4.01	2.50	17.50
10	9.22	3.28	3.13	15.63

Table 2. WER classification results for 20 trials (N = 20).

For this experiment, based on the results from Table 2, the best result was achieved when K = 9, having an average WER of 8.63% and 2.5% WER for the best run. When analyzing the mean WER values across the K values, a clear improvement can be noticed, when the amount of representative examples of each word increases, suggesting that improving the training set might be beneficial for our technique. Additionally, the discrepancy found between the best and worst values suggest that further research is required on how to select the best representation for a certain word.

If we analyze the results from a different perspective, a value stabilization of WER when K = 6 can be observed in the boxplot from Fig. 12. However, considering the available corpora it is important to highlight that when K is higher the amount of test data becomes reduced. For example, when K = 10 only 4 observations from each word are considered. In this graph outliers can also be observed for K = 2, K = 7 and K = 8.

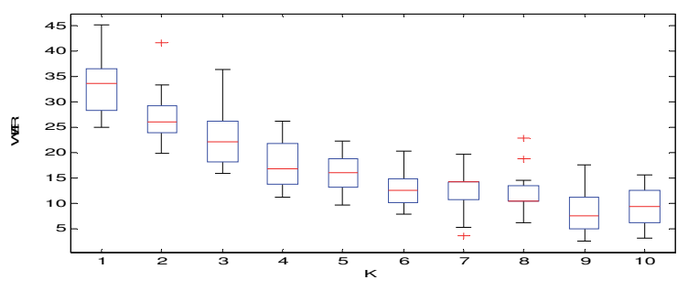


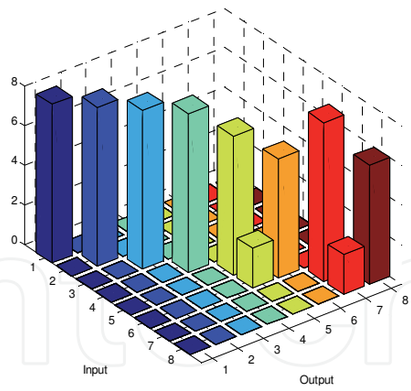
Fig. 12. Boxplot of the WER results for the several K values.

In order to further analyze the quality of the experiment several confusion matrixes are presented (Fig. 13) for the most relevant runs. Each input represents the actual word and each output represents the classified word. The order presented in Table 1 is applied for each word. When analyzing the confusion matrixes for the several trials, errors can be found between the following pairs: [kɐ̃tu] as [katu] (Fig. 13 b)); [matu] as [mɐ̃tu] and vice-versa (Fig. 13 a) and b)); [tĩtu] as [titu] (Fig. 13 a) and b)); and [kɐ̃tu] as [matu] (Fig. 13 c)). As expected, confusion is more often between words where the only difference relies in the nasal sounds (consonants and vowels).

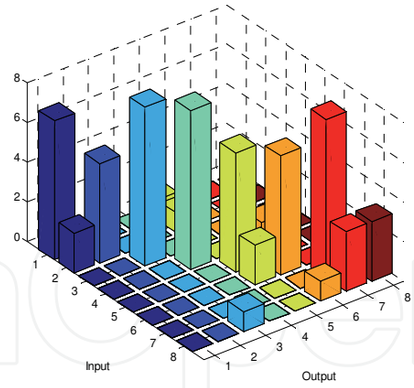
7.2 Experiment II - Doppler-based nasal phonemes detection

In this experiment a Doppler-based sensing (see section 4) system was used, composed by a Doppler emitter connected to a signal generator tuned to 40 kHz and a Dopler receptor tuned to that frequency. Doppler receptor and emitter were placed on both sides of the microphone as described in previous experiments by (Kalgaonkar et al., 2007). Microphone speech signal and Doppler receiver signals were acquired using a 96 kHz sampling rate. The process of demodulation was performed in Matlab and essentially consisted on applying amplitude demodulation to the derivative of the Doppler signal and applying a low pass filter (as in (Kalgaonkar et al., 2007)).

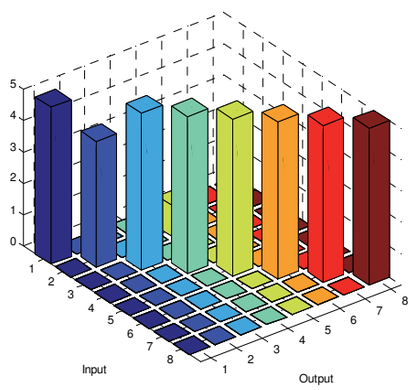
The same set of words used in the previous experiment (see Table 1) was recorded in different conditions by varying distance from speaker to receptor and microphone and silent or normal production. The subject was one of the authors (AT). Below are representative examples of the obtained signals as spectrograms. Fig. 14 presents the results for the minimal pair of words *cato/canto*, essentially differing in the nasality of the first vowel.



a) K = 6 with Best WER = 7.81



b) K = 6 with Worst WER = 20.31



c) K=9 with WER = 2.5%

Fig. 13. Confusion matrix for best and worst run of K = 6 and best run of K = 9. Input and output axis values have the following correspondence with the words from the Corpus: *cato* = 1, *canto* =2, *peta* = 3, *penta* = 4, *mato* = 5, *manto* = 6, *tito* = 7 and *tinto* = 8. The vertical axis corresponds to the number of word classifications.

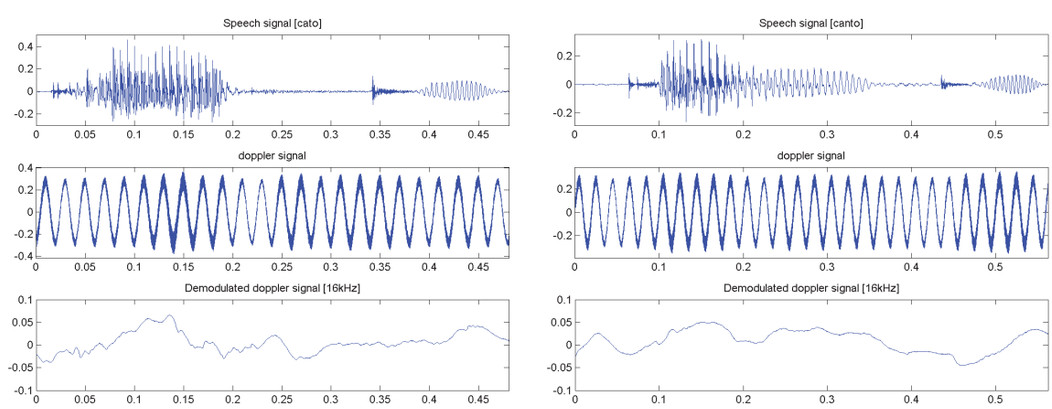


Fig. 14. Speech signal, Doppler signal and demulated signal for the words *cato* and *canto*.

By looking at the speech signal plus the spectral information between 20 and 200 Hz of the demodulated signal in Fig. 15 a clear difference between words can be empirically observed. However, despite the promising look of this approach further research needs to be performed.

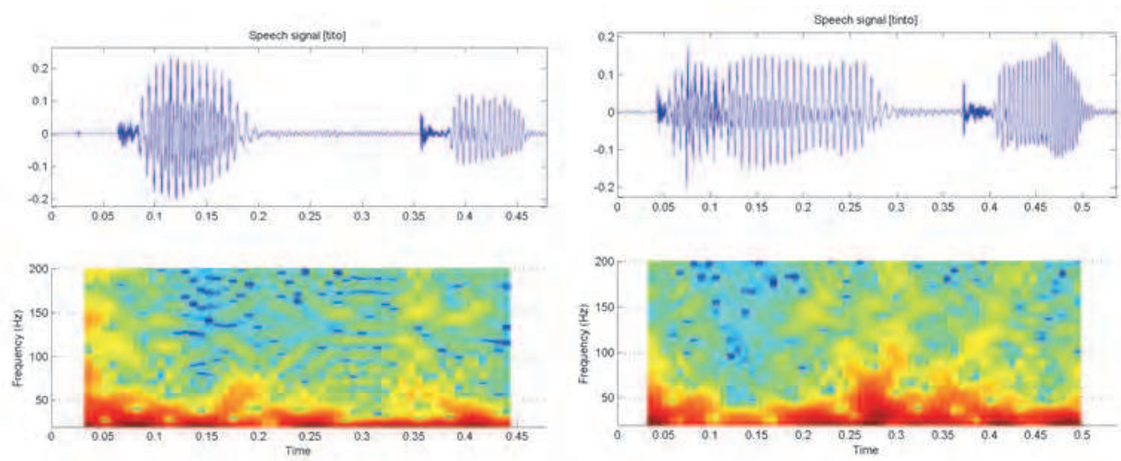


Fig. 15. Speech signal and spectrogram of the demulated signal for the pair *tito/tinto*.

7.2.1 Noticeable differences

This preliminary experiment shows for the first time ADS applied to EP and represents an initial promising step for the development of a SSI based on Doppler for EP. The resulting spectrograms for the several pairs of words show that a clear difference can be subjectively depicted from the obtained spectrograms and time-varying signals, even when the comparison is performed between similar words where the only difference resides in a nasal sound versus an oral one. Further signal processing analysis is needed, to derive new feature vectors towards the improvement of the word classification scheme presented in the previous sections.

8. Conclusion

8.1 Summary of paper and main conclusions from the two experiments

The aim of this work is to identify possible technological solutions and methodologies that enable the development of a multimodal SSI, specially targeted for EP. For that reason a brief analysis of the state-of-the-art was presented and we have proposed a new taxonomy, based on the speech production model. This paper also introduced a novel technique for feature extraction and classification in the VSR domain. This technique is based on an existing robust scale, rotation and luminance invariant feature detection transform in computer vision (FIRST), which has been applied in the past for real-time applications, such as Augmented Reality. The novelty of our approach is the application of FIRST to extract skin features spread across different regions of a speakers' face and to track their displacement, during the time that elapses while the speaker utters a word. By collecting a training set and a test set, totaling 4 minimal pairs of European Portuguese words, with 14 different observations of each word, we were able to classify the silent pronunciation of such words, using a classification scheme based in Dynamic Time Warping (DTW) technique.

DTW was used to find an optimal match between a sufficient number of observations and, in our experiments, we were able to calculate a mean Word Error Rate (WER) of 8.63% (STD 4.01%) with the best figure of 2.5% WER for the best run. As expected, error analysis detected recognition problems between similar words were the only difference is a nasal phoneme instead of an oral one. This demonstrates the difficulty on distinguishing pairs of words that only differ on nasal sounds. However, in this experiment many of these pairs were successfully discriminated, supporting our hypothesis that the skin deformation of the human face, caused by the pronunciation of words, can be captured by studying the local time-varying displacement of skin surface features, using FIRST, and that this framework can be applied to a successful vision-based SSI system for EP. Additionally, first experiments using Doppler signals for EP were reported. In this second experiment, an initial spectral analysis demonstrates a clear difference between the minimal pairs of words mentioned on Table 1. The results from both experiments are promising and motivate the development of a multimodal SSI based on these two technologies.

8.2 Future work

For the VSR modality, we expect to improve the system pipeline, by addressing several approaches. One of such activities is the adoption of the full optimal use of FIRST feature tracking, for real-time silent speech recognition. Additionally, we plan to perform facial segmentation into several regions of interest and the use of a second camera for a lateral perspective of the face, since the experiments carried so far, included only a frontal image from a single camera. We will also consider the use of one or more depth cameras (such as Microsoft Kinect (Kinect, 2011)), along with a parallel research of generalizing FIRST for 3D image features. Regarding the Doppler modality, a complete feature extraction, tracking and classification scheme, similar to the one presented for VSR, needs to be developed and tested. We also plan to carry silent speech recognition experiments in EP involving, the Doppler modality approach. In terms of classification techniques, DTW will be further exploited, but other models that capture the temporal behavior of the signals, such as Hidden Markov Models, could be considered. For statistically significant analysis, we plan to collect and analyze a more extensive corpus in EP, allowing a more concrete investigation of which tracking measures can be extracted during speech. More long term goals include the fusion of both approaches in a single multimodal SSI continuing the studies of how well can both approaches complement one another.

9. References

Bastos, R. & Dias, M. S. (2008). Automatic Camera Pose Initialization, using Scale, Rotation and Luminance Invariant Natural Feature Tracking, in *The Journal of WSCG*

Bastos, R. & Dias, M. S. (2009). FIRST - Fast Invariant to Rotation and Scale Transform: Invariant Image Features for Augmented Reality and Computer Vision. VDM Verlag 2009.

Bay, H.; Tuytelaars, T. & Gool, L. V. (2006). SURF: Speeded Up Robust Features. In *Proceedings of the 9th European Conference on Computer Vision*, Springer LNCS volume 3951, part 1, pp 404-417

Brumberg, J. S.; Kennedy, P. R. & Guenther, F. H. (2009). Artificial speech synthesizer control by brain-computer interface. In *Proceedings of Interspeech 2009*, Brighton, UK.

Brumberg, J. S.; Nieto-Castanonf, A.; Kennedye, P. R. & Guenther. F. H. (2010). Brain–computer interfaces for speech communication. Speech Communication, Volume 52, Issue 4, April 2010, Pages 367-379

Calliess, J.-P. & Schultz, T. (2006). Further Investigations on Unspoken Speech. Studienarbeit, Universita¨ t Karlsruhe (TH), Karlsruhe, Germany

Chan, A.D.C.; Englehart, K.; Hudgins, B.; & 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, 2001, pp. 1727–1730

Chauvenet, W. (1960). A Manual of Spherical and Practical Astronomy Volume II. 1863. Reprint of 1891. 5th ed. Dover, N.Y.: 1960. pp. 474 - 566

DaSalla, C.S.; Kambara, H.; Sato, M. & Koike, Y. (2009). Spatial filtering and single-trial classification of EEG during vowel speech imagery. In: Proc. *3rd Internat. Convention on Rehabilitation Engineering and Q4 Assistive Technology* (i-CREATe 2009), Singapore

Davidson, M. W. & Abramowitz, M. (2006). Molecular Expressions Microscopy Primer: Digital Image Processing - Difference of Gaussians Edge Enhancement Algorithm. In Olympus America Inc. and Florida State University

Denby, B. & Stone, M. (2004). Speech synthesis from real time ultrasound images of the tongue. In: *Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing*, (ICASSP¨04), Montre´al, Canada, 17–21 May 2004, Vol. 1, pp. I685–I688.

Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M. & Brumberg, J.S.(2010), Silent speech interfaces. Speech Communication, v.52 n.4, p.270-287, April, 2010

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

Dupont, S. & Ris, C. (2004). Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise. *In: Proc. Robust 2004, Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, UK, August 2004

Ellis, D. (2003). Dynamic Time Warp (DTW) in Matlab Web resource, last visited on 26-02-2011, available from: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw

Fagan, M.J.; Ell, S.R.; Gilbert, J.M.; Sarrazin, E. & Chapman, P.M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. Med. Eng. Phys. 30 (4), 419–425

Florescu, V-M.; Crevier-Buchman, L.; Denby, B.; Hueber, T.; Colazo-Simon, A.; Pillot-Loiseau, C.; Roussel, P.; Gendrot, C. & Quattrochi, S. (2010). Silent vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface", *Proceedings of Interspeech*, Makuari, Japan

Gurbuz, S., Z. Tufekci, Patterson, E. & Gowdy, J.N. (2001). Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition. *IEEE*

International Conference on Acoustics, Speech, and Signal Processing 2001. Proceedings. (ICASSP '01),

Harris, C. & Stephens, M. (1988). A combined corner and edge detector. *In Proceedings of the 4th Alvey Vision Conference*, pp 147-151

Helfrich, H. (1979). Age markers in speech. In: Scherer, K. & Giles, H.: Social markers in speech. Cambridge: University Press

Heracleous, P.; Nakajima, Y.; Lee, A.; Saruwatari, H. & Shikano, K. (2003). Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation. *Automatic Speech Recognition and Understanding, 2003*. ASRU '03. 2003 IEEE Workshop on , vol., no., pp. 73- 76, 30 Nov.-3 Dec. 2003

Hueber, T.; Aversano, G.; Chollet, G.; Denby, B.; Dreyfus, G.; Oussar, Y.; Roussel, P. & Stone, M. (2007). Eigentongue feature extraction for an ultrasound-based silent speech interface. Proceedings of ICASSP (Honolulu, USA), pp. 1245-1248

Hueber, T.; Chollet, G.; Denby, B.; Dreyfus, G. & Stone, M. (2008). Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface. Interspeech, Brisbane, Australia, pp. 2032-2035

Hueber T.; Benaroya, E. L.; Chollet, G.; Denby, B.; Dreyfus, G.; Stone, M. (2009). Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface. *In Proceedings of Interspeech 2009*, Brighton, UK

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

Jou, S.; Schultz, T. & Waibel, A. (2004). Adaptation for Soft Whisper Recognition Using a Throat Microphone. International Conference of Spoken Language Processing (ICSLP-2004), Jeju Island, South Korea, October 2004.

Jou, S.; Schultz, T. & Waibel, A. (2007). Multi-stream articulatory feature classifiers for surface electromyographic continuous speech recognition. In: Internat. Conf. on Acoustics, Speech, and Signal Processing. IEEE, Honolulu, Hawaii

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

Kalgaonkar K. & Raj B. (2008). Ultrasonic doppler sensor for speaker recognition," in ICASSP'08

Ke, Y. & Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR), pp 506-513

Kinect (2011). Last visited on 27-02-2011, Available from: http://www.xbox.com/pt-PT/kinect

Lacerda, A. & Head, B. F. (1966). Análise de sons nasais e sons nasalizados do Português. Revista do Laboratório de Fonética Experimental (de Coimbra), VI:5_70.

Levelt, W. (1989). Speaking: from Intention to Articulation. Cambridge, Mass.: MIT Press.

Liang, Y.; Yao, W.; Minghui, D. (2010). Feature Extraction Based on LSDA for Lipreading. International Conference on Multimedia Technology (ICMT)

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication* 22(1): 1-15

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision 60*, pp. 91-110

MacGurk, H. & MacDonald, J. (1976). *Hearing lips and seeing voices*, Nature, Vol. 264, pp. 746–748

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

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

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

Nakajima, Y.; Kashioka, H.; Shikano, K. & Campbell, N. (2003). Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin, In: Proc. IEEE ICASSP, pp. 708–711

Nakajima, Y. (2005). Development and evaluation of soft silicone NAM microphone. Technical Report IEICE, SP2005-7, pp. 7–12

Ng, L.; Burnett, G.; Holzrichter, J. & Gable, T. (2000). Denoising of human speech using combined acoustic and EM sensor signal processing. In: *Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 5–9 June 2000, Vol. 1, pp. 229–232

Nijholt, A.; Tan, D.; Pfurtscheller, G.; Brunner, C.; R. Millán, J.; Allison, B.; Graimann, B.; Popescu, F.; Blankertz, B. & Müller, K. (2008). Brain-Computer Interfacing for Intelligent Systems, IEEE Intelligent Systems, Vol.23 No.3, pp.72-79, (May 2008)

Otani, M.; Shimizu, S. & Tatsuya, H. (2008). Vocal tract shapes of non-audible murmur production, Acoustical Science and Technology, 29(2): 195-198

Patil, S. A. & Hansen, J. H. L. (2010). The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification, Speech Communication 52(4): 327-340

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

Porbadnigk, A.; Wester, M.; Calliess, J. & Schultz, T. (2009). EEG-based speech recognition impact of temporal effects, In: Biosignals 2009, Porto, Portugal, (January 2009), pp. 376–381

Potamianos, G.; Neti, C.; Gravier, G.; Garg, A. & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech, Proceedings of the IEEE 91(9): 1306-1326

Quatieri, T.F.; Messing, D.; Brady, K.; Campbell, W.B.; Campbell, J.P.; Brandstein, M.; Weinstein, C.J.; Tardelli, J.D. & Gatewood, P.D. (2006). Exploiting non-acoustic sensors for speech enhancement, IEEE Trans. Audio Speech Lang. Process, 14 (2), 533–544

Rabiner, L. R. & Juang, B. (1993). Fundamentals of speech recognition, Prentice-Hall, Inc., Chapter 4

Rossato, S.; Teixeira, A. & 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, FR, Jun. 2006.

Sá, F.; Afonso, P.; Ferreira, R. & Pera, V. (October 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, October 8, 2003

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

Shi, J. & Tomasi, C. (1994). Good Features to Track, In IEEE Conference on CVPR

Srinivasan, S.; Raj, B. & Ezzat, T. (2010). Ultrasonic sensing for robust speech recognition, In ICASSP

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

Stork, D. G. & Hennecke, M. E. (1996). Eds., Speechreading by Humans and Machines, Berlin, Germany: Springer-Verlag

Suppes, P.; Lu, ZL. & Han, B. (1997). Brain wave recognition of words, Proc Natl Acad Sci USA 94:14965–14969

Teixeira, A. & 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), pp. 183-193, ICMC-USP, Atibaia, São Paulo, Brasil

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

Teixeira, A.; Martinez, R.; Silva, L. N.; Jesus, L..; Príncipe, J. C.; Vaz, F. (2005). Simulation of Human Speech Production Applied to the Study and Synthesis of European Portuguese. *Eurasip Journal on Applied Signal Processing*. Hindawi Publishing Corporation, vol. 2005, nº 9, p. 1435-1448

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

Tran, V.-A.; Bailly, G.; Loevenbruck, H. & Toda, T. (2010). Improvement to a NAM-captured whisper-to-speech system, Speech Communication, Vol.52, Issue 4, (April 2010), pp.314-326

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

Toda T.; Nakamura K.; Nagai T.; Kaino T.; Nakajima Y. & Shikano, K. (2009). Technologies for Processing Body-Conducted Speech Detected with Non-Audible Murmur Microphone. *In Proceedings of Interspeech 2009*, Brighton, UK; September 2009.

Toth, A.R.; Kalgaonkar, K.; Raj, B. & Ezzat, T. (2010). Synthesizing speech from Doppler signals. *IEEE International Conference* on *Acoustics Speech and Signal Processing* (ICASSP), vol., no., pp.4638-4641, 14-19 March 2010

Wand, M.; Jou, S.; Toth, A. R. & Schultz, T. (2009). Synthesizing Speech from Electromyography using Voice Transformation Techniques. *In Proceedings of Interspeech 2009*, Brighton, UK

Wester, M. & Schultz, T. (2006). Unspoken speech – speech recognition based on electroencephalography. Master's Thesis, Universita¨t Karlsruhe (TH), Karlsruhe, Germany

Zhao, G.; Barnard, M. & Pietikäinen, M. (2009). Lipreading with local spatiotemporal descriptors. Trans. Multi. 11(7): 1254-1265

Zhu, B.; Hazen, T.; J. & Glass, J. R. (2007). Multimodal speech recognition with ultrasonic sensors. In Eurospeech

Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7

Hard cover, 432 pages

Publisher InTech

Published online 23, June, 2011

Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Joao Freitas, Antonio Teixeira, Miguel Dias and Carlos Bastos (2011). Towards a Multimodal Silent Speech Interface for European Portuguese, Speech Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, Available from: http://www.intechopen.com/books/speech-technologies/towards-a-multimodal-silent-speech-interface-for-european-portuguese

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821