Enhancing Multimodal Silent Speech Interfaces with Feature Selection

João Freitas[1,2], *Artur Ferreira*[3,4], *Mário Figueiredo*[3,5], *António Teixeira*[2], *Miguel Sales Dias*[1,6]

[1] Microsoft Language Development Center, Lisboa, Portugal
[2] Dep. Electronics Telecommunications & Informatics/IEETA, University of Aveiro, Portugal
[3] Institute of Telecommunications (IT), Lisboa, Portugal
[4] Dep. Elect., Tele. & Computers, Lisbon Engineering Institute (ISEL), Lisboa, Portugal
[5] Dep. Electronics and Computers, Instituto Superior Técnico (IST), Lisboa, Portugal
[6] ISCTE - University Institute of Lisbon (ISCTE-IUL), Lisboa, Portugal

t-joaof@microsoft.com, arturj@isel.pt, mtf@lx.it.pt, ajst@ua.pt, miguel.dias@microsoft.com

Abstract

In research on Silent Speech Interfaces (SSI), different sources of information (modalities) have been combined, aiming at obtaining better performance than the individual modalities. However, when combining these modalities, the dimensionality of the feature space rapidly increases, yielding the well-known “curse of dimensionality”. As a consequence, in order to extract useful information from this data, one has to resort to feature selection (FS) techniques to lower the dimensionality of the learning space. In this paper, we assess the impact of FS techniques for silent speech data, in a dataset with 4 non-invasive and promising modalities, namely: video, depth, ultrasonic Doppler sensing, and surface electromyography. We consider two supervised (mutual information and Fisher’s ratio) and two unsupervised (mean-median and arithmetic mean geometric mean) FS filters. The evaluation was made by assessing the classification accuracy (word recognition error) of three well-known classifiers (k-nearest neighbors, support vector machines, and dynamic time warping). The key results of this study show that both unsupervised and supervised FS techniques improve on the classification accuracy on both individual and combined modalities. For instance, on the video component, we attain relative performance gains of 36.2% in error rates. FS is also useful as pre-processing for feature fusion.

Index Terms: Multimodal Silent Speech Interface, Feature Selection, Supervised Classification.

1. Introduction

Silent Speech Interfaces (SSI) allow for speech communication with a system in the absence of an acoustic signal [1]. By analyzing data gathered during different parts of the human speech production process, these interfaces allow users with speech impairments (e.g. users who have been subject to a partial or total laryngectomy) to communicate with a system. SSI can also be used in the presence of environmental noise, in situations in which privacy, confidentiality, or non-disturbance are important. Some SSI techniques based on different sensor types have been proposed in SSI-related literature, namely: ElectroMagnetic Articulography (EMA) sensors [2], UltraSound (US) jointly with optical imaging of the tongue and lips [3], Ultrasonic Doppler Sensing (UDS) [4], and surface ElectroMyoGraphy (EMG) [5], among others. However, some of these techniques are either highly invasive or require clinical intervention, making them unusable as a natural interface (e.g. implants in the speech-motor cortex [6]). When considering non-invasive approaches, such as UDS, the reported accuracy rates in a digit recognition task [4] are inferior to those found in conventional automatic speech recognition. A system that could explore the advantage of the strongest points of a particular modality would help to mitigate the weaknesses of the individual modalities. Nowadays, there are a few datasets for SSI research, particularly for multimodal scenarios. When multiple input modalities are considered, the large dimensionality of the feature space augments the complexity of the recognition task. To address these problems, in the literature of SSI we find many approaches that rely on Feature Reduction (FR) techniques, such as Linear Discriminant Analysis (LDA) [5][7]. However, the use of FR techniques such as LDA do not allow us to unveil and to interpret directly the modalities and/or features that are more relevant for the task at hand. FR techniques generate a new set of features, which are functions of the original ones that correspond to the physical process. Thus, for SSI data it may be preferable to apply a Feature Selection (FS) method in order to filter and keep a subset of the original features. Moreover, it has been found that many FR techniques such as LDA, may not perform well with low amounts of data for High-Dimensional (HD) spaces [8][9].

Our aim is to develop a multimodal SSI that combines information from different parts of the human speech production process [10] and fulfills the following requirements: possibility of being used in a natural manner without complex medical procedures, low cost, tolerant to noisy environments and able to work with speech-handicapped users. Given these requirements, a novel type of SSI based on the following specifications was defined as our target: (1) facial information acquired from Visual and Depth sensors; (2) surface EMG of the articulator muscles; (3) capture of facial movements during speech using UDS. After synchronously acquiring data from these modalities, our aim is to understand which information stream achieves the best results in a speech recognition task and how the existing redundancy, among these modalities, affect the overall results.

The rest of this paper is organized as follows: Section 2 summarizes recent work on multimodal SSI and presents a brief overview of FS techniques. Section 3 describes the methods applied in this study, namely the system used to acquire multiple streams of data, the features extracted from each stream, the FS techniques that we have considered, and the evaluated classifiers. Section 4 reports the experimental results for individual modalities and as well as for their fusion, with and without FS. A discussion of the results is provided in section 5. Finally, section 6 ends the paper with some concluding remarks.

2. Background and Related Work

2.1. Multimodal Silent Speech Interfaces

Regarding multimodal silent speech, in 2004, Denby and Stone [11] reported a seminal experiment in which 2 input modalities, in addition to the acoustic signal, were used to develop a SSI. Denby and Stone applied ultrasound imaging of the tongue area along with video information of the lips. In 2008, Tran et al. [12] also reported a preliminary approach using information from 2 modalities: whispered speech acquired using a Non-Audible Murmur (NAM) microphone and visual information of the face using the 3D position of 142 colored beads glued to the speakers face. Later, using the same modalities, these authors reported an absolute improvement of 13.2% when adding the visual information to the NAM data stream. The use of visual facial information combined with surface EMG signals has also been proposed by Yau et al. in 2008 [13]. In their study, Yau et al. present a SSI that analyses the possibility of using surface EMG for unvoiced vowels recognition and a vision-based technique for consonant recognition. Recently, in 2010, Florescu et al. [3] reported a 65.3% recognition rate using the same modalities in an isolated word recognition scenario with a 50-word vocabulary and a Dynamic Time Warping (DTW) based classifier.

In 2013, some of the authors, presented a multimodal SSI database and a preliminary classification experiment that combined Video, Depth, surface EMG and UDS [14]. In this experiment the best result for an isolated digit recognition task was found for Video plus Depth with 72.1% error rate.

2.2. Feature Selection in multimodal SSI

Feature Selection (FS) techniques aim at finding adequate subsets of features for a given learning task [12][13]. The use of FS techniques may improve the accuracy of a classifier learnt from data by helping to avoid the so-called “curse of dimensionality” and may speed up the training time while improving the generalization processes. In a broad sense, FS techniques are classically grouped into four main types of approach: wrapper, embedded, filter, and hybrid methods [15][16][17]. Among these four types, filter approaches are characterized by assessing the adequacy of a given subset of features solely using characteristics of the data, without resorting to any learning algorithm or optimization procedure. It is often the case that for HD data, such as SSI data, the filter approach is the only one that produces acceptable results in terms of their running-time [18]. For this reason, despite the different types of approaches, in this paper we consider solely filter FS methods. There are decades of research on FS, for different problems (e.g. [15][16][17]). However, in the context of multimodal signal processing, the research for FS methods has received little attention. Recently, an approach based on information theory, for audio–visual speech recognition, the which checks for redundancy among features, yielding better performance than LDA, has been proposed [19];

3. Methodology and Tools

The adopted method consists to apply FS techniques to a new multimodal database and to assess the recognition results in an isolated-word recognition task. This section provides details on the SSI database, the features extracted, the employed FS techniques as well as the classifiers applied on the selected data.

3.1. Multimodal Database: Acquisition and Data

The data of the modalities was acquired using: (1) a Microsoft Kinect for Windows that acquires visual and depth information; (2) a surface EMG sensor acquisition system from Plux [20] with 5 pairs of EMG surface electrodes, which captures the myoelectric signal from the facial muscles; (3) a custom built dedicated circuit board [21] (referred hereon as UDS device) based on the work of Zhu et al. [22]. Details about the device connections, the synchronization process, positioning of the sensors and the acquired data can be found in [23].

We have selected a vocabulary of 10 European Portuguese digits, from zero to nine. The corpus was recorded by three native speakers (one male and two female) with 31, 65, and 71 years old. No history of hearing or speech disorders is known for these speakers. Each speaker has recorded six repetitions per word, yielding a total of 180 utterances. Each utterance was pronounced individually, in a random order.

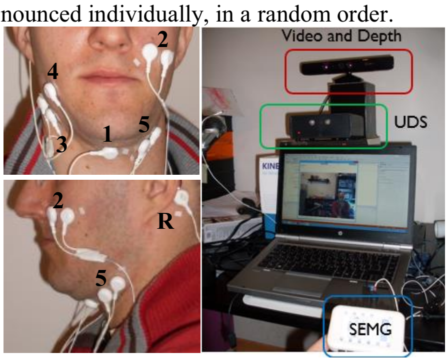

Figure 1: *Left - EMG electrodes positioning and their channels (1 to 5) plus the reference electrode (R). Right – the acquisition setup with all its devices.*

3.2. Feature Extraction

In this study, we have chosen the most recent Feature Extraction (FE) techniques for each modality as well as those that reported the best results.

For Video and Depth, we started by establishing a Region of Interest (ROI) containing the lips and surrounding areas. Using real-time Active Appearance Models (AAM) [24], we were able to obtain a 64x64 pixel ROI centered at the speaker’s mouth. Then, we apply an appearance based method, which, due to variations in illumination, skin color, facial hair and other factors, are usually preferred to shape based methods. In this context, one of the most common approaches is to use a Discrete Cosine Transform (DCT) [25] in the ROI. Following previous studies, we compressed the pixel information by computing the DCT of the 64x64 block and keeping the low spatial frequencies by selecting the first 64 coefficients contained in the upper left corner of the coefficients matrix. We have only considered the odd columns of the DCT, in order to take advantage of the facial symmetry and imposing horizontal symmetry to the image. After applying the 2D DCT, the first and second temporal derivatives are appended to the feature vector, generating a final feature vector of 192 dimensions per frame. The variation between speakers and recording conditions are smoothed by using Feature Mean Normalization (FMN) [26].

In terms of FE, we have followed a similar approach to that of [4] and started by pre-processing the UDS signal. The acquired signal was first zero-averaged, then a third order moving average filter was applied to suppress the carrier and finally a difference operator was applied. After the pre-processing stage, we split the signal into 50ms frames with a 10ms frame shift, and applied a 2048 points Discrete Fourier Transform (DFT) [25] to the preprocessed signal for the bandwidth around the carrier, 2275 Hz to 5200 Hz. Finally, the DCT is applied to the DFT coefficients to decorrelate and compress the signal, extracting the first 38 coefficients. The signal energy, velocity, and acceleration coefficients were appended to the DCT coefficients, yielding a final feature vector of 117 dimensions per frame.

For surface EMG feature extraction, we used an approach which is based on temporal features, similar to the one described in [5] as well as to previous work from the authors [27], (without FR). The features were extracted for each EMG signal frame of 30ms. A frame shift of 10ms was considered. A context width of 15 frames was also used, generating a final feature vector of 155 dimensions per channel. Finally, we stacked all the channels in a single feature vector of 775 dimensions.

3.3. Feature Selection Filter

In this work, we consider two unsupervised and two supervised relevance measures. For the unsupervised case, we consider: i) the Mean-Median (MM), that is, the absolute value of the difference between the mean and the median of a feature (an asymmetry measure) [28]; ii) the quotient between the Arithmetic Mean and the Geometric Mean (AMGM) of each feature, after exponentiation (a dispersion measure) [28]. We also consider two well-known supervised measures: i) the (Shannon's) Mutual Information (MI) [29], which measures the dependency between two random variables; ii) the Fisher's ratio [30], which measures the dispersion among classes.

For finding the most relevant features, we have considered the Relevance-Redundancy FS (RRFS) method proposed in [28]. In a nutshell, RRFS uses a relevance measure to sort the features in decreasing order, and then performs a redundancy elimination procedure on the most relevant ones. At the end, it keeps the most relevant features exhibiting up to some Maximum Similarity (MS) between themselves. The similarity between features is assessed with the Absolute Cosine (AC) of the angle between feature vectors, say X_i and X_j, given by

$$AC_{ij} = |\cos(\theta_{ij})| = \left| \frac{< X_i, X_j >}{\|X_i\| \|X_j\|} \right| \qquad (1)$$

where <.,.> denotes inner product between vectors and ||.|| is the L2 norm of a vector. AC yields 0 for orthogonal feature vectors and 1 for collinear ones. RRFS has been applied successfully to other types of high-dimensional data [28].

3.4. Classifiers and Evaluation

For classification and in order to explore different techniques we have considered the k-Nearest Neighbor (kNN) [31], Support Vector Machines (SVM) [32][33], and DTW classifiers [34]. The kNN classifier uses cosine distance for prediction and the *k* number of neighbors is dynamically determined, as the square root of *n* (the number of training instances) [35]. The SVM classifier from LIBSVM [36] was configured with a linear kernel. The DTW classifier uses the distance between the test word samples and each sample of the training set choosing the class with the minimum distance.

We split our dataset into train and test using a stratified 9-fold strategy, splitting the database into 160 training utterances and 20 test utterances (2 test utterances per class) in each run. The error rate herein reported is estimated from the average error rate of the 9 folds.

4. Experimental Results

In order to assess the benefits of applying FS techniques to this dataset, we start by estimating the recognition error of each modality without applying FS methods, using the most common features of each modality, to establish a baseline recognition error. Afterwards, we apply FS techniques to each individual modality and to multiple modalities, (fusion scenario).

4.1. Baseline Results - without Feature Selection

Table 1 shows the error rate for the baseline using *all* the features (see section 3.2). Among the four modalities, the best result was attained by surface EMG, with 46.7%, followed by UDS with 50.6%. Depth information presents the worst result, with 70.6%. The SVM classifier attains the best average accuracy, 55.5%, followed by the kNN classifier with 64.9%. For comparison purposes, we have also applied FR using LDA to this dataset. In the achieved results only Depth information results with the SVM classifier improves accuracy with a relative performance gain of 23.7%. The remaining results are either similar (e.g. Video using kNN) or are worse.

Table 1. *Baseline results (without FS). Average word recognition error rate (%) with 95% confidence interval (9-fold), for each modality.*

Classifier	Video	Depth	EMG	UDS
kNN	74.4±6.2	71.7±6.1	52.2±6.1	57.8±4.7
SVM	53.9±6.3	70.6±5.8	46.7±4.9	50.6±5.3
DTW	66.7±1.6	73.9±4.3	83.9±2.7	72.8±4.7

4.2. Single Modality Analysis - Feature Selection

In this section, we address the results of each individual modality, using each relevance criterion (see section 3.3) for FS. Table 2 reports the best results of kNN, SVM, and DTW, after applying RRFS and considering MS values between 0.1 and 0.9.

Table 2. *Average word recognition error rate (%) with 95% confidence interval (9-fold), per classifier for each FS technique. The rightmost column is the relative improvement of the best result with respect to the baseline.*

Mod. / Class.		Mutual Info	Fishers ratio	MM	AMGM	Rel. Imp.
Video	kNN	67.2±4.4	66.7±5.4	63.3±4.0	66.7±6.5	14.9%
	SVM	41.1±4.8	**34.4±5.3**	42.8±7.1	40.6±4.5	36.2%
	DTW	48.9±6.9	43.9±5.1	65.0±4.6	65.0±4.9	34.2%
Depth	kNN	70.6±7.4	70.6±6.8	68.9±6.1	70.6±6.4	3.9%
	SVM	70.6±4.1	**64.4±6.6**	68.3±3.7	67.2±7.5	8.8%
	DTW	72.2±4.0	70.0±6.5	75.6±4.1	72.2±5.4	5.3%
EMG	kNN	51.7±5.9	50.0±6.9	52.2±7.5	55.0±5.7	4.2%
	SVM	46.7±6.5	**45.0±3.7**	46.1±5.1	46.1±3.6	3.6%
	DTW	68.9±3.6	82.2±4.7	72.8±3.7	67.8±6.3	19.2%
UDS	kNN	56.1±4.3	56.1±4.8	56.1±4.6	55.6±6.4	3.8%
	SVM	50.6±5.3	50.0±6.1	**47.8±5.2**	50.6±5.3	5.5%
	DTW	76.7±3.7	67.8±4.4	71.7±4.0	77.8±4.4	6.9%

The use of a FS technique usually improves on the results, for each individual modality, as compared to the baseline (Table 1), with a few exceptions. The best results were achieved using SVM. Comparing all techniques with the baseline, we attain an average of 12.2% relative improvement. The best results with FS techniques were found for Video with an average 28.4% improvement for the 3 classifiers. Regarding FS, the best results were achieved by the supervised Fisher's ratio for Video, Depth, and EMG modalities; using the unsupervised MM, for UDS we get the best result of 47.8%.

The MS parameter values for the best results seem to vary according to the modality, the FS technique, and the classifier. For instance, for Video using the SVM classifier, we have the largest improvement with MS=0.3. In the second best case, for EMG using SVM, MS=0.8.

In terms of reduction (between the resulting subset and the original one), the best result was achieved for Video with a reduction of 95.1% of the original features, followed by Depth with 59.8%. Surface EMG achieved a reduction of 26.7% whereas UDS attains only 5%.

4.3. Multiple Modality Analysis - Feature Selection

We now assess the combination of multiple modalities. When fusing the feature vectors of each modality a slight improvement can be noticed, as compared to the results of Table 1. For example, Video combined with EMG improves the results of Video by 14.5% and the results of EMG with 2.4%. Similar improvements can be noticed for UDS when combining it with the EMG information.

When considering RRFS before two modality combinations, improvements can be noticed for almost all combinations when using SVM. The best results were achieved for the case of Video combined with UDS using Fisher's ratio and Video combined with Depth using AMGM, with relative performance gains of 36.1%, and 19.2%, respectively. Regarding the combination of three and four modalities, we notice an interesting pattern in which the previous results are somewhat replicated. For instance, adding UDS to the combination of Video+Depth shows the same resultsas Video+Depth. However, adding EMG (and consequently including all streams) only improves the baseline results found in Table 1. Table 3 reports the most relevant results for the SVM classifier (the one that reported, the best results in this analysis).

The overall results show that an average improvement of 15.3% is achieved across all possible combinations of two modalities and 6.0% for combinations of three and four modalities.

Table 3. *Average word recognition error rate with 95% confidence interval (9-fold), before and after FS (best result) using SVM. The rightmost column is the relative improvement with respect to the baseline.*

Modalities	Best results	After FS	Rel. Imp.
Video+Depth	69.4±6.8	56.1±5.1 (AMGM)	19.2%
Video+UDS	53.7±6.1	34.3±5.2 (Fisher)	36.1%
Video+EMG	46.1±4.6	45.0±3.7 (Fisher)	2.4%
EMG+UDS	46.7±4.9	45.0± 3.7 (Fisher)	3.6%
Video+Depth+UDS	69.4±6.8	56.1±5.1 (AMGM)	19.2%
All Modalities	46.1±4.6	46.1±4.6 (AMGM)	0%

5. Discussion

When comparing FR and FS techniques, FS presents better results. The most probable cause is the small amount of training data relative to the sample dimensionality. Previous studies [8][9][19] have shown that when this situation occurs, the LDA within-scatter matrix becomes sparse, reducing the efficiency of the LDA transform. In terms of individual modalities, we have found that FS techniques improve the results of all considered modalities, with a noticeable improvement in Video and with Depth producing the worst results, in accordance to what was found previously in the literature of Audio-Visual Speech Recognition [7][19].

Regarding the considered FS techniques, on average, supervised techniques performed better than unsupervised ones; the best results were achieved by Fisher's ratio on three modalities; for the UDS modality, MM achieved the best result. We have also analyzed the combination of modalities and the effect of FS, before feature fusion. In this case, we have noticed the following: (1) if no FS is applied, small improvements can still be noticed, particularly for the worst individual modalities; (2) after applying FS, similar levels of error rates can be achieved to the ones attained individually, with noticeable improvements for the case of Depth input.

Selecting the best features can also be used as a way to compress data, removing redundant information and making it appropriate for scenarios where data storage or communication bandwidths are reduced (e.g. video modality under mobile communication).

6. Conclusions

This paper assesses the impact of feature selection on silent speech data. We report a study in which feature selection filters with relevance-redundancy assessment are applied to the data. In detail, two supervised (Mutual Information and Fisher's ratio) and two unsupervised (Mean-Median and Arithmetic Mean Geometric Mean) relevance measures are applied to several non-invasive and promising Silent Speech Interfaces modalities (Video and Depth input, Surface Electromyography, and Ultrasonic Doppler). The attained results show that feature selection leads to improvements, which can be achieved either using only a single modality or a combination of several modalities. Looking at the modalities, the highest improvement was found for Video alone and for Video combined with UDS using an SVM classifier. In terms of feature selection techniques, the supervised methods based on Fisher's ratio attained the best results. The results of this study also show that the feature vectors which are currently accepted as the state-of-the-art can be reduced in average 46.6% (considering all modalities) and up to the maximum of 95.1% for the Video component. As future work, we would like to explore and analyze the combination of FS with feature discretization techniques.

7. Acknowledgements

This work was partially funded by Marie Curie IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP), Marie Curie Golem (ref.251415, FP7-PEOPLE-2009-IAPP) and by National Funds (FCT- Foundation for Science and Technology) in the context of IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEst-C/EEI/UI0127/2011).

8. References

[1] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., and Brumberg, J.S, "Silent speech interfaces", *Speech Communication*, 52(4):270-287, April 2009.

[2] Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E. and Chapman, P.M, "Development of a (silent) speech recognition system for patients following laryngectomy", *Med. Eng. Phys*., 30(4):419–425, 2008.

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

[4] Srinivasan, S., Raj, B. and Ezzat, T., "Ultrasonic sensing for robust speech recognition", *Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 5102-5105, 2010.

[5] Schultz, T. and Wand, M., "Modeling coarticulation in large vocabulary EMG-based speech recognition". *Speech Communication*, 52(4):341-353, April 2010.

[6] Brumberg, J.S., Nieto-Castanonf, A., Kennedye, P.R. and Guenther, F.H., "Brain–computer interfaces for speech communication", *Speech Communication*, 52(4):367-379, April 2010.

[7] Galatas, G., Potamianos, G. and Makedon, F., "Audio-visual speech recognition incorporating facial depth information captured by the Kinect," *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2714-2717, August 2012.

[8] Qiao, Z., Zhou, L., and Huang, J.Z., "Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data". *International Journal of Applied Mathematics*, 39:48–60, 2009.

[9] Bickel, P. and Levina E., "Some theory for Fishers linear discriminant function naive Bayes, and some alternatives when there are many more variables than observations", *Bernoulli*, 10(6):989-1010, 2004.

[10] Levelt, W., "Speaking: from Intention to Articulation", *Cambridge, Mass.: MIT Press*, 1989.

[11] Denby, B., Stone, M., "Speech synthesis from real time ultrasound images of the tongue", *Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, 1: I685–I688, May 2004.

[12] Tran, V.A., Bailly, G., Loevenbruck, H. and Jutten, C., "Improvement to a NAM captured whisper-to-speech system", *Proceedings of Interspeech 2008*, 1465-1468, 2008.

[13] Yau, W.C., Arjunan, S.P. and Kumar, D.K., "Classification of voiceless speech using facial muscle activity and vision based techniques", *TENCON 2008 IEEE Region 10 Conference*, 2008.

[14] Freitas J., Teixeira, A., Dias, M. S., "Multimodal Silent Speech Interface based on Video, Depth, Surface Electromyography and Ultrasonic Doppler: Data Collection and First Recognition Results". *Int. Workshop on Speech Production in Automatic Speech Recognition*, Lyon, 2013.

[15] Guyon, I. and Elisseeff. A., "An introduction to variable and feature selection". *Journal of Machine Learning Research (JMLR)*, 3, 1157–1182, 2003.

[16] Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. (Editors). Feature extraction, foundations and applications. Springer, 2006.

[17] Das, S., "Filters, wrappers and a boosting-based hybrid for feature selection", *Proceedings of the International Conference on Machine Learning (ICML)*, 74–81, 1994.

[18] Yu, L. and Liu., H., "Feature selection for high-dimensional data: a fast correlation based filter solution". *Proceedings of the International Conference on Machine Learning (ICML)*, 856–863, 2003.

[19] Gurban, M., Thiran, J-P., "Information Theoretic Feature Extraction for Audio-Visual Speech Recognition", *IEEE Transactions on Signal Processing*, 57(12):4765-4776, 2009.

[20] Plux Wireless Biosignals, Portugal, Online: http://www.plux.info/, accessed on 25 March 2014.

[21] Freitas, J. Teixeira, A., Vaz, F. and Dias, M.S., "Automatic Speech Recognition based on Ultrasonic Doppler Sensing for European Portuguese", *Advances in Speech and Language Technologies for Iberian Languages*, 328:227-236, Springer, 2012.

[22] Zhu, B., Hazen, T.J. and Glass, J., "Multimodal speech recognition with ultrasonic sensors", *Proceedings of Interspeech, 2007*. Antwerp, Belgium, 2007.

[23] Freitas, J., Teixeira, A., Dias, M.S., "Multimodal Corpora for Silent Speech Interaction", *Proceedings of 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May 2014.

[24] Cootes, T.F., Gareth J.E., and Christopher J.T., "Active appearance models", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(6):681-685, 2001.

[25] A. Oppenheim, R. Schafer, Discrete-Time Signal Processing, Prentice Hall, 2009.

[26] Potamianos, G., Neti, C., Gravier, G., Garg A. and Senior, A.W., "Recent advances in the automatic recognition of audiovisual speech", *Proceedings of the IEEE*, 91(9):1306-1326, 2003.

[27] Freitas, J., Teixeira, A. and Dias, M.S., "Towards a Silent Speech Interface for Portuguese: Surface Electromyography and the nasality challenge", *Internat.. Conf. on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*, Vilamoura, Portugal, 2012.

[28] Ferreira, A. and Figueiredo, M., "Efficient Feature Selection Filters for High-Dimensional Data", *Pattern Recognition Letters*, 33(13):1794-1804, October 2012

[29] Cover, T. and Thomas, J., "Elements of Information Theory", *John Wiley & Sons*, 1991.

[30] Fisher, R., "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, 7:179–188, 1936.

[31] Aha, D. W., Kibler, D. and Albert, M. K. "Instance-based learning algorithms". Machine learning, 6(1):37-66, 1991.

[32] Vapnik, V., "The nature of statistical learning theory". Springer-Verlag, 2000.

[33] Burges. C., "A tutorial on support vector machines for pattern recognition". *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[34] D. Ellis (2003). Dynamic Time Warp (DTW) in Matlab. Online: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/, accessed on 25 March 2014.

[35] De Wachter, M., "Example based continuous speech recognition," *Ph.D. thesis*, K. U. Leuven, ESAT, Belgium, 2007.

[36] Chang C.-C. and Lin, C.-J., "LIBSVM : a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology*, 2(3):1-27, 2011.