32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Syllable-Based Speech Recognition Using EMG

Eduardo Lopez-Larraz, Oscar M. Mozos, Javier M. Antelis, Javier Minguez

Abstract— **This paper presents a silent-speech interface based on electromyographic (EMG) signals recorded in the facial muscles. The distinctive feature of this system is that it is based on the recognition of syllables instead of phonemes or words, which is a compromise between both approaches with advantages as (a) clear delimitation and identification inside a word, and (b) reduced set of classification groups. This system transforms the EMG signals into robust-in-time feature vectors and uses them to train a boosting classifier. Experimental results demonstrated the effectiveness of our approach in three subjects, providing a mean classification rate of almost 70% (among 30 syllables).**

I. INTRODUCTION

The most natural and powerful way of communication for humans is the spoken language. For this reason there has been vast research in learning the design principles of systems able to understand human speech and expressions. Natural language communication with machines is typically done using automatic speech recognition (ASR) systems. The usual setting is a user that speaks to a microphone, and then, the ASR recognizes the speech and the integrated application behaves according to the established dialogue.

One of the main drawbacks of traditional speech interfaces is their limited robustness in the presence of ambient noise [1], [2]. To overcome this limitation, several electromyographic (EMG) approaches have been proposed in which the acoustic speech recognition is substituted by silent-speech recognition. The classification is based on the myoelectric signals produced in the facial muscles during speech [3], [4], [5], [6], [7], [8]. This solution overcomes the ambient noise but also provides an alternative to human-machine communication for people with speech disabilities such as laryngectomy, as well as elderly or convalescent people. In these cases there is no acoustic signal coming from the user, or the signal is distorted or very weak.

Focusing in existing and natural EMG speech recognition systems, there are mainly three possible approaches to the problem. The first one is based in phoneme recognition. This problem has about 30 classes (approximately the number of letters in Spanish language) but the main difficulty is to delimit where begins and finishes a phoneme inside a word. Examples of these systems and limitations can be seen in [4], [5], [6] (vowel recognition) and [7]. Another possibility is to use complete words recognition [1], [2], [8], [9]. Here, the difficulty emerges as the number of words to classify increases dramatically (they cover the full language). Thus existing systems reduce the search to a very limited set of words. This paper proposes a halfway solution to obtain a natural speech recognizer, which is the recognition of the syllables (common and natural way to divide words in Spanish language). Syllables are easier to identify inside a word than phonemes due to the fact that they are simply voice hits and correspond to abrupt muscle movements. Additionally, they are several orders of magnitude smaller in number than words (a person uses about 3000 words in his/her day by day speech), so it provides a trade off between the two opposite methods.

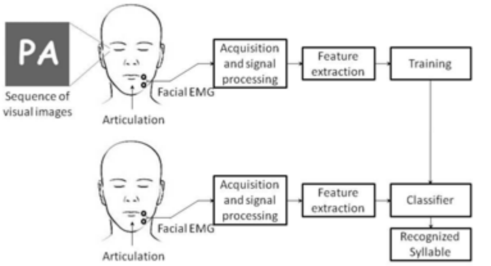

Fig. 1. The image shows the two steps of our EMG-based speech recognition system: calibration (top sequence), and online operation (bottom sequence).

This paper presents a silent-speech interface based on EMG signals recorded from the facial muscles. The prosthesis prototype is displayed in Fig. 1 where it has a calibration step and then an online step. The calibration step uses EMG signals recorded during speech to train a classifier. The second step classifies the online EMG to detect the intended syllable of the user. As a previous step, this paper proposes a study of applicability based on offline recognition of a representative number of Spanish syllables and different classification strategies.

II. METHODS

This section describes the methodology followed in the work: the definition of the vocabulary, the anatomic location and instrumentation issues of the EMG system, and the feature extraction and classification.

A. Definition of the Vocabulary

The objective of the present recognition system is to recognize isolated syllables from the Spanish language, which are usually composed by a consonant followed by a vowel. All the syllables are divided into five groups according to the

E. Lopez-Larraz, O.M. Mozos, J.M. Antelis and J. Minguez are with the Instituto de Investigación en Ingeniería de Aragón (I3A) and Dpto. de Informática e Ingeniería de Sistemas (DIIS), Universidad de Zaragoza, Spain. Email: {edulop,ommozos,antelis,jminguez}@unizar.es. This work has been partially supported by projects HYPER-CSD2009- 00067, DPI2009-14732-C02-01 funded by the Spanish Government.

978-1-4244-4124-2/10/$25.00 ©2010 IEEE

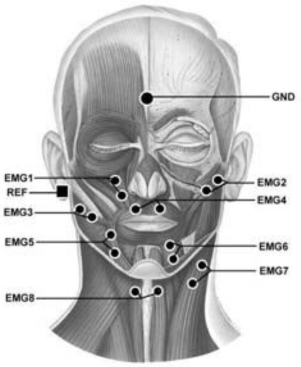

Fig. 2. Selected facial muscles for EMG recording: *Levator labii superioris* (EMG1) , *Zygomaticus major* (EMG2), *Risorius* (EMG3), *Orbicularis oris* (EMG4), *Depressor anguli oris* (EMG5), *Depressor labii inferioris* (EMG6), *Platysma* (EMG7) and *Anterior belly of the digastric* (EMG8).

anatomical articulation origin [10]: labials, dentals, palatals, velars, and alveolars. In order to have a representative set of syllables for the system that cover all groups, it was selected one representative consonant of each group combined with the five vowels plus the vowels separately. The final set was composed by 30 syllables, presented in Table I.

TABLE I
COMPLETE SET OF SYLLABLES

Vowels	/a/	/e/	/i/	/o/	/u/
Labials	/pa/	/pe/	/pi/	/po/	/pu/
Dentals	/ta/	/te/	/ti/	/to/	/tu/
Palatals	/ya/	/ye/	/yi/	/yo/	/yu/
Velars	/ka/	/ke/	/ki/	/ko/	/ku/
Alveolars	/la/	/le/	/li/	/lo/	/lu/

B. Facial Electromyography for Speech Recognition

Electromyography signals reflect the electrical activity of the muscles during a movement. In the case of speech, EMG signals are generated in the facial muscles responsible for pursing the lips, lifting the corners of the mouth, or opening the jaw. Additionally, EMG signals also appear in the extrinsic muscles of the tongue, which are responsible for relaxing the tongue up and forward. According to previous anatomical studies [11], the number of muscles involved in speech production is very high. This makes the recording of all possible facial EMG-signals almost impractical. Furthermore, there exists no standard selection of the most appropriate muscles for EMG-based speech recognition, thus this selection is typically done in a heuristic way.

In this work, the EMG electrodes were placed on the facial muscles according to their distinctive movements during the pronunciation and articulation of speech utterances in Spanish [10]. To reduce the posterior complexity of the system, EMG electrodes were placed on muscles only on one side of the face, since they are symmetric. The final muscles selection and EMG locations are shown in Fig. 2.

C. Facial EMG Recording

The preparation of the EMG electrodes followed the guidelines proposed in [12]. Face skin areas over the site of the facial muscles were previously cleaned with alcohol-wetted swabs. Conductive electrode gel was added to the electrodes to minimize the impedance at the skin-electrode surface contact. Bipolar electrodes were placed in the same direction of the fibers of the facial muscle and the distance was fixed to be 1 cm. The ground electrode (GND in Fig. 2) was placed on the forehead, and the reference electrode (REF in Fig. 2) was placed on the left earlobe. The impedance at each electrode was checked to be below 10 kΩ. The eight bipolar EMG signals were acquired and digitized (using a gUSBamp amplifier from gTec) at a sampling frequency of 2400 Hz, power-line notch-filtered to remove the 50 Hz line interference, and band-pass filtered between 5 and 500 Hz to remove different noise sources out of the EMG signals frequency band. The general instrumentation was a commercial gTec amplifier, and eighteen gold-made EMG surface electrodes (diameter: 10 mm). The recording system and software was developed under the BCI2000 platform [13].

D. Experimental Protocol and Data Collection

Three healthy male students of our university with no known speech impediments or disorders, and whose native language is Spanish participated in the experiments. The participants were duly informed about the whole protocol of the study. In the experimental recording sessions, the participants were sitting in front of a computer screen and the EMG sensors were placed over the skin surface according to Fig. 2. In each experimental session, the EMG signals corresponding to 50 examples of each of the Spanish syllables listed in Table I were recorded, yielding a total of 1500 collected examples per subject. One session was divided in 75 trials (the inter-trial time was 60 seconds), where in each trial 20 syllables were randomly shown to the subject. During the execution of a trial, a dark screen was first displayed during 10 seconds to rest the participant before the visual stimuli presentation. Subsequently, an image with the required syllable was shown and the participants articulated the syllable without producing voice. Each image was displayed during 1 second, and followed by a grey screen of 1 second in order to relax the facial muscles before the next syllable.

In order to test the robustness of the classifier across time, one subject repeated a session articulating only the vowels. Each vowel was repeated 50 times following the previous protocol.

E. EMG Feature Extraction and Classification

The recorded EMG data was used to train a classifier for these syllables. There are two steps in this process: the feature extraction and the classification. On the one hand, the signal representation is the set of features selected to represent the raw signals (examples) and to train the classifier. A relevant selection criterion is that they have to be robust to time shifts. This is because the user does not pronounce all the syllables at the same point in time. Thus,

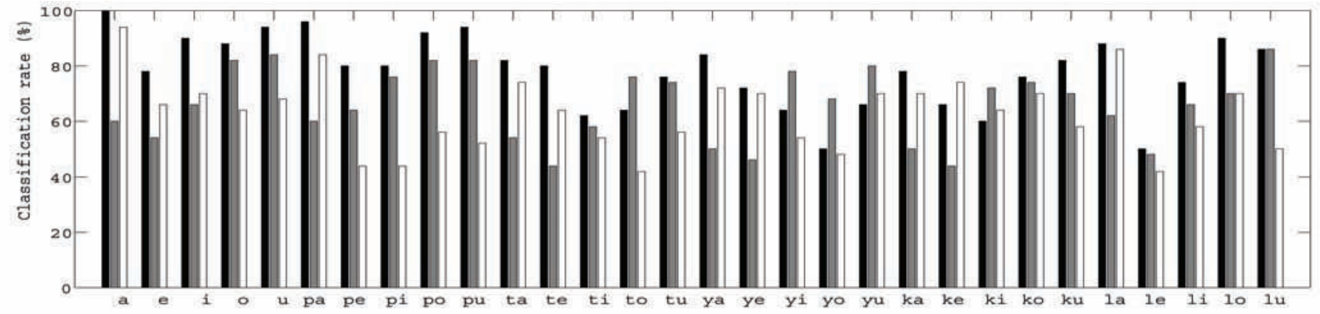

Fig. 3. True positive classification rates for each of the three subjects.

all the selected features were time-shift invariant to have a classification process not sensitive to the pronunciation time. The following features were selected according to previous works in silent speech recognition [1], [4], [5], [7], [8], [9]: Fast Fourier Transform (20 components), Root Mean Square, Average amplitude of the signal, Maximum amplitude, Kurtosis, Mel-frequency cepstral coefficients (13 values, as in [8]), Mean absolute value, Zero-crossing points, Sum of all the signal values and Sum of all the rectified signal values. Each raw signal of an EMG channel is transformed into a feature vector whose 41 components are the characteristics listed before. The final feature vector of the complete signal (representing one syllable example) is obtained by the concatenation of the different features vectors for each channel. Notice that the dimension of the final feature vector is much lower than the dimension of the original signal ($41 \times 8 \ll 2400 \times 8$), which is a significant reduction in complexity of the problem without loss in the classification results (as we will see in the experiments).

On the other hand, a classifier is trained to distinguish the examples (represented by the previous feature vectors) from the different syllables. The classifier selected was the boosting algorithm AdaBoost.M1 [14], using the J4.8 decision tree [15] as the weak classifier. This tree is a variation of the C4.5 decision tree [16] and applies a post-pruning method to improve the classification performance of the final model. The training and classification processes was carried out using the software Weka [17].

III. EXPERIMENTS

This section describes the classification results of the 30 syllables using an offline 10-fold cross validation, a validation in different sessions, and an evaluation of different classifiers.

A. Results with the Complete Set of Syllables

The first experiment validated the recognition system individually for each subject using 10-fold cross validation. The true positive classification rates for the three subjects are depicted in Fig. 3 (a true positive is a syllable that has been correctly classified). The mean recognition rates for the three subjects were 78.07%, 66.00% and 62.94% respectively, and the global mean rate was 69%. Additionally, all the classes obtained accuracies higher than 40% (notice that a random classifier would provide a 3.33%). Furthermore, the confusion matrix of the mean classification results for the three subjects is shown in Fig. 4. Notice that the most salient confusions in the classification form groups (see Section II-A and Table I), in which the initial phoneme is the same (e.g. a clear group of confusions appeared among the syllables starting with /y*/), or in terminations with the same letter (e.g. /ta/ and /ya/, /te/ and /ye/...). These behaviors are not that bad because there are clear patterns behind the confusions (the most frequent confusions for one classification inside a group is restricted most of the time to four or five possibilities). In general, these results suggest a high performance and potential of the recognition system given the large number of classes involved in the problem.

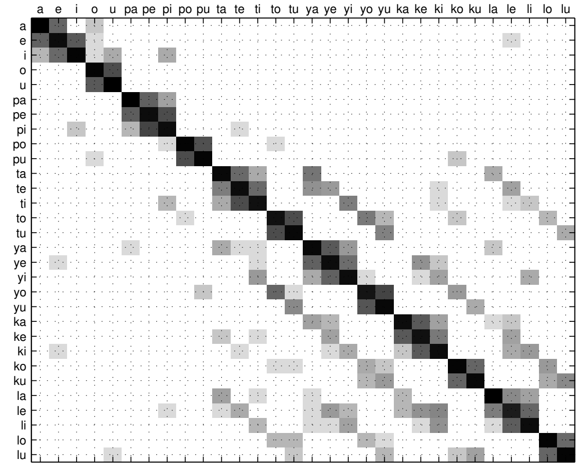

Fig. 4. Confusion matrix for the classification of the 30 syllables. Dark values indicate a high classification rate. To increase the visibility, we used a logarithmic scale. The classifications rates are calculated using the mean of the classification rates for the three subjects.

B. Training Data from Different Sessions

The second issue is to study the influence of the data recorded in different days. Table II shows the performance of the classifier using the articulated vowels of the first session only. Then, the 50 recorded vowels of the first session were added to the 50 vowels of a second session to train the classifier. The performance of the new resulting classifier is shown in Table III, where can be noticed that the performance is highly similar. This indicates that, for the same user, the trained system has a certain degree of invariance with respect to different points in time and possible small variations in the electrodes placement.

TABLE II

CONFUSION MATRIX FOR THE CLASSIFICATION OF THE VOWELS IN THE FIRST SESSION

	/a/	/e/	/i/	/o/	/u/
/a/	**90%**	10%	0%	0%	0%
/e/	8%	**60%**	30%	0%	2%
/i/	0%	22%	**76%**	0%	2%
/o/	2%	0%	0%	**84%**	14%
/u/	0%	0%	0%	14%	**86%**

TABLE III

CONFUSION MATRIX FOR THE CLASSIFICATION OF THE COMBINED VOWELS FROM TWO DIFFERENT SESSIONS

	/a/	/e/	/i/	/o/	/u/
/a/	**92%**	4%	4%	0%	0%
/e/	5%	**71%**	23%	1%	0%
/i/	3%	24%	**72%**	0%	1%
/o/	0%	1%	0%	**85%**	14%
/u/	0%	2%	1%	19%	**78%**

C. Comparison of Different Learning Approaches

The last issue is to study the application of four different strategies to create the final classifier. They are formed by the combination of two possible representations of the examples and two learning approaches. The signals could be represented by the original raw sampled data or by the computed feature vector introduced in Sect. II-E. Besides, the classifier could be in one case the J4.8 decision tree, and in the other a combination of Adaboost.M1 and J4.8 decision trees (cf. Sect. II-E). The resulting performances are shown in Table IV.

TABLE IV

CORRECTLY CLASSIFIED INSTANCES GIVEN BY TWO DIFFERENT ALGORITHMS AND TWO DIFFERENT SIGNAL REPRESENTATIONS.

	Decision Tree	AdaBoost + Decision Tree
Raw sampled signal	42.2%	61.9%
Computed Features	69.5%	**80.2%**

According to Table IV, the use of feature vectors improves the performance of the algorithms in approximately 20% over the raw sampled signals. Moreover, the use of boosting augments the classification rates in more than 10%. These results justify the selection of feature vectors to represent the signals, and the boosting process to improve the final performance of the decision trees.

IV. CONCLUSIONS AND FUTURE WORKS

This paper presents a prototype of a silent-speech recognition system based on electromyographic signals recorded in facial muscles. The approach focussed on syllables of the Spanish language. The signals from each articulated syllable were transformed into a feature vector whose components represented different global characteristics. A classifier based on boosting was trained using these feature vectors as input. Experiments carried out with three different subjects demonstrated the effectiveness of the proposed system when recognizing new articulated syllables.

The future work focuses on the usage of the current syllable classifier as basis to build a recognition system of complete Spanish words. The interest is in using techniques to classify sequences of observations such as hidden Markov models.

REFERENCES

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

[2] A. D. C. Chan, K. Englehart, B. Hudgins, and D. F. Lovely, "Hidden Markov Model Classification of Myoelectric Signals in Speech," in *Proceedings of the Annual Conference of the IEEE/EMBS*, Istanbul, Turkey, October 2001.

[3] C.-N. Huang, C.-H. Chen, and H.-Y. Chung, "The Review of Applications and Measurements in Facial Electromyography," *Journal of Medical and Biological Engineering*, vol. 25, no. 1, pp. 15–20, 2004.

[4] S. P. Arjunan, H. Weghorn, D. K. Kumar, and W. C. Yau, "Vowel recognition of English and German language using Facial movement (SEMG) for Speech control based HCI," in *Proceedings of the HCSNet Workshop on Use of Vision in Human-Computer Interaction*, Canberra, Australia, 2006.

[5] S. P. Arjunan, D. K. Kumar, W. C. Yau, and H. Weghorn, "Unspoken Vowel Recognition Using Facial Electromyogram," in *Proceedings of the IEEE EMBS Annual International Conference*, New York City, USA, August 2006.

[6] J. A. Mendes, R. R. Robson, S. Labidi, and A. K. Barros, "Subvocal Speech Recognition Based on EMG signal Using Independent Component Analysis and Neural Network MLP," in *Proceedings of the Congress on Image and Signal Processing*, Hainan, China, 2008.

[7] Q. Zhou, N. Jiang, K. Englehart, and B. Hudgins, "Improved Phoneme-Based Myoelectric Speech Recognition," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2016–2023, 2009.

[8] H. Manabe and Z. Zhang, "Multi-stream HMM for EMG-Based Speech Recognition," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, USA, September 2004.

[9] C. Jorgensen, D. D. Lee, and S. Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," *Proceedings of the International Joint Conference on Neural Networks*, January 2003.

[10] A. R. Mestre, *La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico*, 4th ed. Laboratorio de Lingüística Informática de la Universidad Autónoma de Barcelona, 1998, (in Spanish).

[11] B. Tuller, K. S. Harris, and B. Gross, "Electromyographic study of jaw muscles during speech," *Journal of Phonetics*, vol. 9, pp. 175–188, 1981.

[12] A. J. Fridlund and J. T. Cacioppo, "Guidelines for Human Electromyographic Research," *Psychophysiology*, vol. 23, no. 5, pp. 567–589, 1986.

[13] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1034–1043, June 2004.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the European Conference on Computational Learning Theory*, 1995, pp. 23–37.

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

[16] R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[17] Weka, "Weka 3: Data Mining Software in Java," http://www.cs.waikato.ac.nz/ml/weka/.