Syllable-Based Speech Recognition Using EMG

Eduardo Lopez-Larraz, Oscar M. Mozos, Javier M. Antelis, Javier Minguez.
I3A and University of Zaragoza, Spain, 2010.

Introduction

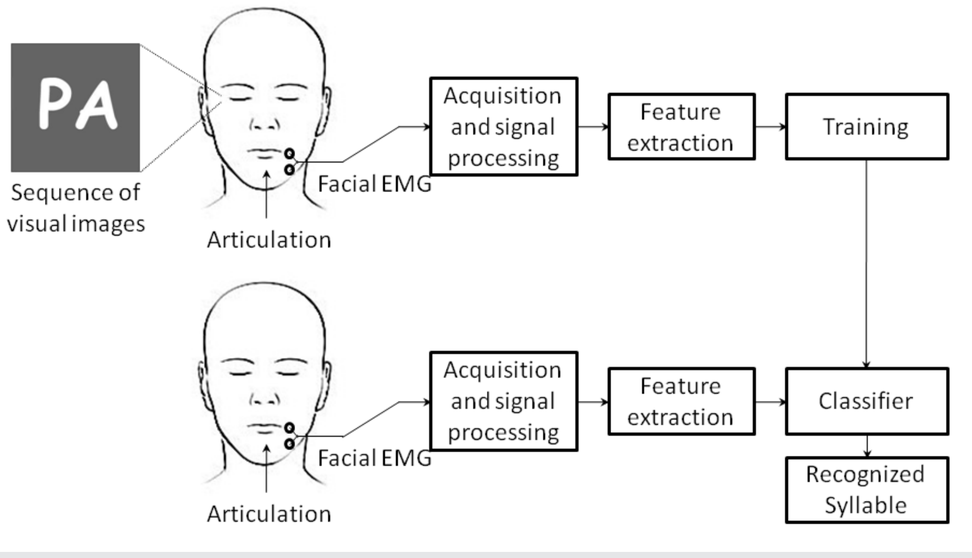
Problem Statement

To design a prototype able to **acquire facial EMG signals** produced during pronunciation of Spanish syllables, **transform them into feature vectors**, and feed a classifier which will **recognize the syllable performed.**

To perform experiments with different subjects in ordert to **test the effectiveness** of the prototype **when recognizing** (i) **a high number of syllables** and (ii) **syllables acquired in different experimentation sessions.**

Solution

Problems of signal classification are typically divided into two steps: (a) calibration (top sequence) and (b) online operation (bottom sequence). Step (a) consists in acquire as many examples of each class as possible to train a classifier, which in step (b) will identify new executions of the trained classes. In this problem the classes to recognize are EMG signals recorded from the facial muscles during articulation of different Spanish syllables.



Benefits

These protoypes are being designed in order to complement automatic speech recognition (ASR) systems in noisy environments, where their performance decreases dramatically, and as a way of communication for people with speech disabilities, such as laryngectomy [1].

Methods

Definition of the Vocabulary

A set of 30 most representative and used syllables from the Spanish language was chosen:

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/

Experiments

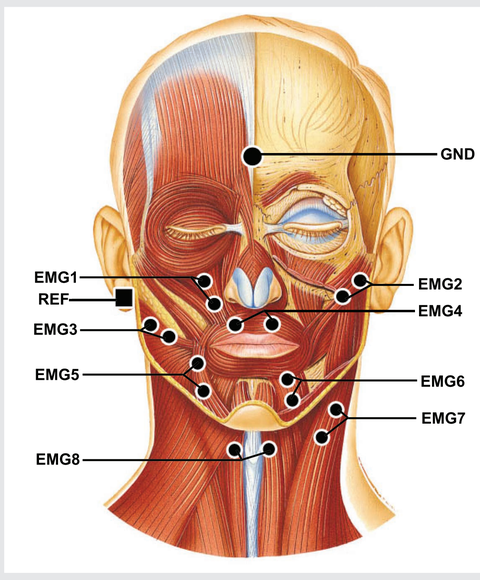
Three healthy subjects participated in the experimental sessions.

For each participant were recorded 50 samples per syllable.

One of the subjects repeated a session articulating 50 times the vowels only.

EMG-Electrodes Location

Bipolar electrodes were placed on 8 facial muscles to record the EMG produced during pronunciation of the syllables:



EMG1: Levator labii superioris
EMG2: Zygomaticus major
EMG3: Risorius
EMG4: Orbicularis oris
EMG5: Depressor anguli oris
EMG6: Depressor labii inferioris
EMG7: Platysma
EMG8: Anterior belly of the digastric
GND was placed on the forehead
REF located on the right earlobe

Feature Extraction

The features must be robusts to time shifts to have a classification process not sensitive to the pronunciation time.

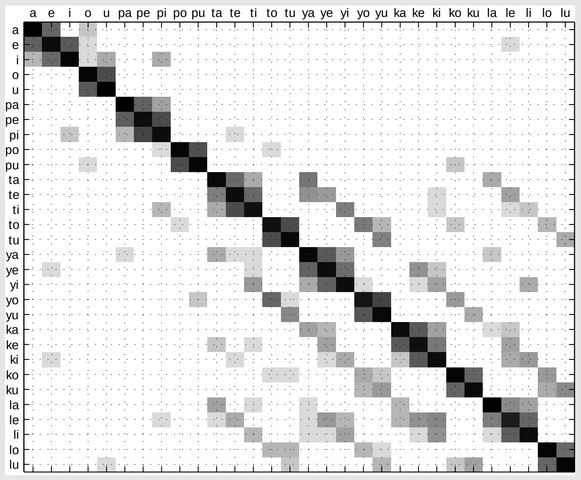
For each channel were extracted 42 coefficients. The final feature vector of each syllable was formed by the concatenation of the corresponding vectors of each channel.

The transformations applied were the following:

- Fast Fourier Transform (20 coeffs.)
- Root Mean Square
- Average amplitude of the signal
- Maximum amplitude
- Mel-frecquency cepstral coefficients (13 coeffs.)
- Sum of the signal values
- Sum of the rectified signal values
- Kurtosis
- Mean absolute value
- Zero-crossing points

Results

Percentages of the Complete Set of Syllables



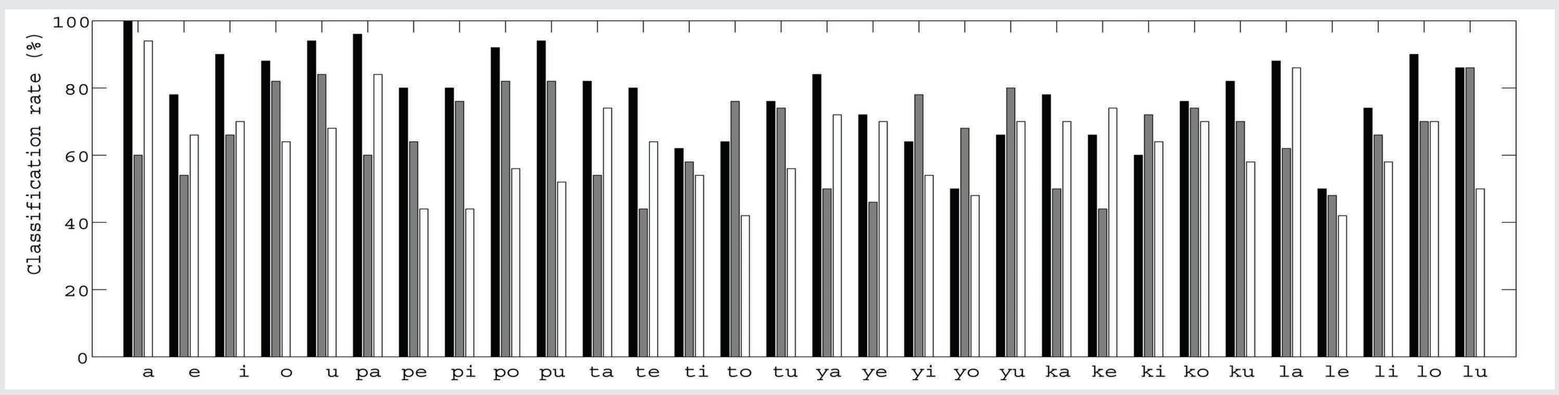
Left figure shows a confusion matrix with the classification rates for the 30 syllables.

The values were calculated averaging the mean coefficients for the three subjects. Dark values indicate a high rate.

Notice that he most salient confusions form groups, where the initial or the last phoneme is the same.

Bottom figure displays the true positive classification rates for each of the three subjects. All the classes have accuracies higher than 40%, while a random classifier would produce a 3.33%.

The average rate for the three subjects is 69%.



Training Data from Different Sessions

Performance of the classifier using only the articulated vowels of the first session for one of the subjects.

	/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** %

Results produced by the classifier when adding to the previous data the 50 extra examples of each vowel recorded in the second session of that participant.

	/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** %

[1] S. P. Arjunan, D. K. Kumar, W. C. Yau, and H. Weghorn, "Unspoken Vowel Recognition Using Facial Electromyogram", IEEE EMBS Annual International Conference, New York City, 2006.