Expert Systems With Applications 66 (2016) 42–61

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multimodal emotion recognition with evolutionary computation for human-robot interaction

Luis-Alberto Perez-Gaspar, Santiago-Omar Caballero-Morales*, Felipe Trujillo-Romero

Technological University of the Mixteca, Road to Acatlima K.m. 2.5, Huajuapan de León, Oaxaca, Mexico,69000, Mexico

ARTICLE INFO

Article history:
Received 4 December 2015
Revised 12 August 2016
Accepted 13 August 2016
Available online 3 September 2016

Keywords:
Emotion recognition
Principal Component Analysis
Hidden Markov Models
Genetic Algorithms
Artificial Neural Networks
Finite state machines

ABSTRACT

Service robotics is an important field of research for the development of assistive technologies. Particularly, humanoid robots will play an increasing and important role in our society. More natural assistive interaction with humanoid robots can be achieved if the emotional aspect is considered. However emotion recognition is one of the most challenging topics in pattern recognition and improved intelligent techniques have to be developed to accomplish this goal. Recent research has addressed the emotion recognition problem with techniques such as Artificial Neural Networks (ANNs)/Hidden Markov Models (HMMs) and reliability of proposed approaches has been assessed (in most cases) with standard databases. In this work we (1) explored on the implications of using standard databases for assessment of emotion recognition techniques, (2) extended on the evolutionary optimization of ANNs and HMMs for the development of a multimodal emotion recognition system, (3) set the guidelines for the development of emotional databases of speech and facial expressions, (4) rules were set for phonetic transcription of Mexican speech, and (5) evaluated the suitability of the multimodal system within the context of spoken dialogue between a humanoid robot and human users. The development of intelligent systems for emotion recognition can be improved by the findings of the present work: (a) emotion recognition depends on the structure of the database sub-sets used for training and testing, and it also depends on the type of technique used for recognition where a specific emotion can be highly recognized by a specific technique, (b) optimization of HMMs led to a Bakis structure which is more suitable for acoustic modeling of emotion-specific vowels while optimization of ANNs led to a more suitable ANN structure for recognition of facial expressions, (c) some emotions can be better recognized based on speech patterns instead of visual patterns, and (d) the weighted integration of the multimodal emotion recognition system optimized with these observations can achieve a recognition rate up to 97.00 % in live dialogue tests with a humanoid robot.

1. Introduction

Human perception regarding emotion detection is a natural process that involves social interactions. Technological advances are in constant evolution and the future aims to a life with robotic agents capable of understanding people. Technology is growing very fast and research on the development of service robots for elderly people or people with motor disabilities (Odashima et al., 2008; PAL Robotics, 2015) is being performed and reported on the literature.

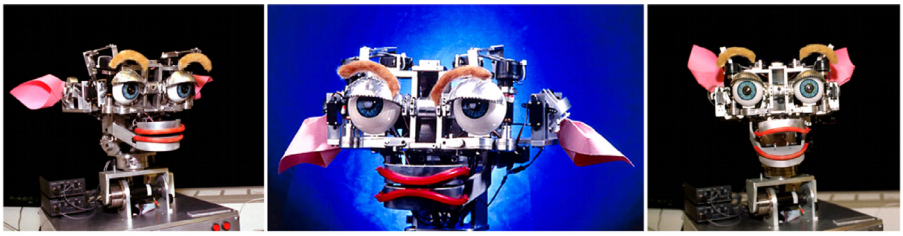
In recent years, research on the affective relation between a robot and a human has become an important subject. Developments on artificial intelligence have been focused on trying to emulate how humans interact with each other. Emotions are fundamental for social interaction and a robot with this capability can take the role of a companion entity that can support conversation, understanding, and responses aimed to improve the well-being of the human user. Emotion recognition can lead to the development of systems for more natural, understandable and intuitive communication (Samani & Saadatian, 2012).

Among the robotic systems that integrate emotion recognition for an interaction task, the following can be mentioned:

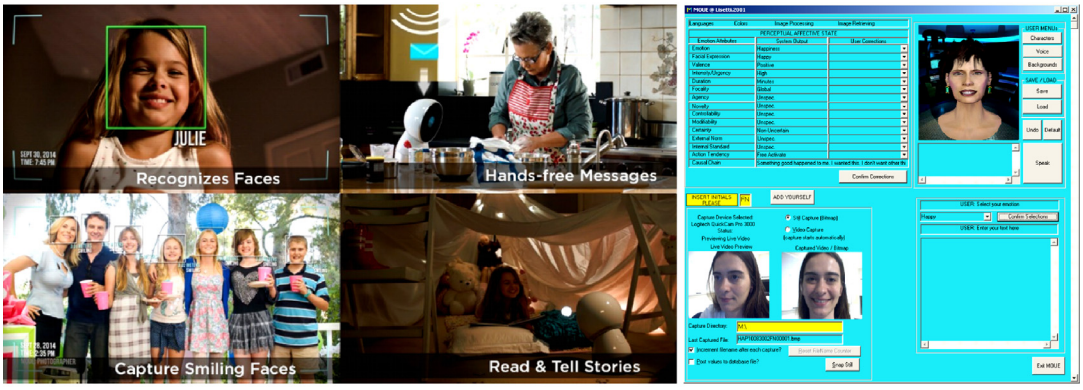
- Robot Kismet (Breazeal, 2003): This social robot was developed by Cynthia Breazeal at the Massachusetts Institute of Technology (MIT). This robot was developed to study how emotions expressed by a robotic system changed the perception and interaction with human users. Kismet was able to recognize affective intentions through the voice (Anger, Fear, Happiness, Tiredness,

* Corresponding author.
E-mail addresses: luis_335450@hotmail.com (L.-A. Perez-Gaspar), scaballero1979@yahoo.com (S.-O. Caballero-Morales), ftrujillo@mixteco.utm.mx (F. Trujillo-Romero).

http://dx.doi.org/10.1016/j.eswa.2016.08.047
0957-4174/



(a) Kismet



(b) Jibo (c) MOUE

Fig. 1. Examples of artificial systems with emotion recognition.

Disgust, Surprise, Sadness, Interest and Calm) and express different emotions as presented in Fig. 1(a).

- Jibo (Chambers et al., 2015): This social robot was an extension of Kismet and it was conceived as a companion robot for commercial purposes. As presented in Fig. 1(b) Jibo *could* perform different tasks like taking photographs, reminding important events, telling stories and establishing video conference between relatives using Wi-Fi.
- Model Of User's Emotions (MOUE) (Lisetti et al., 2003): This was an intelligent interface for distance patient monitoring. This system was able to capture physiological signals and emotional gestures through a bracelet connected to a computer and a web-cam. Then this data was collected and sent to a central computer for processing. As presented in Fig. 1(c) this system enabled question-based interaction with an animated character (avatar) that acted as a mirror by reflecting the facial expressions performed by the patient. The interface was able to process the emotions of Neutral, Anger, Fear, Sadness and Frustration.

The development of robotic systems with the capability of emotion recognition depends on the use and adaptation of pattern recognition techniques. Among the most common techniques the following can be mentioned: Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Principal Component Analysis (PCA), Fuzzy Logic, Hidden Markov Models (HMMs) and Linear Discriminant Analysis (LDA). In Tables 1 and 2 a review of works on emotion recognition is presented. This review includes information regarding the number of emotions that were detected, the patterns (voice, facial expressions) that were considered, and the recognition techniques that were applied.

Further improvement on emotion recognition can be achieved by the development of multimodal systems that integrate expressions and voice patterns. In (Song et al., 2008) a multimodal system was built with Tripled HMMs (THMMs) for the recognition of the following emotions: Surprise, Happiness, Anger, Fear, Sadness and Neutral. The use of THMMs was performed to synchronize voice and facial features in the time domain. For the vision system, the distances between eyes, eye-nose, mouth-nose and width of the mouth were considered as the most important features. For the speech system, 48 prosodic and 16 formant features were extracted. The recognition rates were 87.40% and 81.45% for the vision and speech systems respectively. However, when an integration of both systems was realized an increase in the recognition rate was accomplished, leading to a final recognition rate of 93.31%. Similar improvements were reported in (Busso et al., 2004; Haq et al., 2008) for the recognition rate of four and seven emotions respectively. In Table 3 a review of works on the development of multimodal systems is presented.

From the works presented in Tables 1, 2 and 3 some observations can be highlighted:

- In general, the recognition rates of the vision systems are higher than those of the speech systems (Anagnostopoulos et al., 2015). However a multimodal system can achieve higher recognition rates than those of the individual vision or speech systems (Busso et al., 2004; Song et al., 2008).
- Most of the emotion recognition works on facial expressions considered specialized databases like FEEDTUM (Filko & Martinovic, 2013; Pal & Hasan, 2014), JAFFE (Gosavi & Khot, 2013; Kaur, Vashisht, & Neeru, 2010; Pooja & Kaur, 2010; Rasoulzadeh, 2012; Thuseethan & Kuhanesan, 2014), FACES (Tayal & Vijay, 2012), CK+ (Chaturvedi & Tripathi, 2014), and RaFD (Ilbeygi & Hosseini, 2012). These works also reported the highest emotion recognition rates.
- While most of the emotion recognition works on facial expressions consider six emotions (Chaturvedi & Tripathi, 2014; Gosavi & Khot, 2013; Ilbeygi & Hosseini, 2012; Pal & Hasan, 2014; Rasoulzadeh, 2012; Thuseethan & Kuhanesan, 2014), most of the works on speech recognition consider four (Caballero, 2013; Firoz-Shah et al., 2009; Lee et al., 2004; Wu & Liang, 2011; Yu et al., 2001) and five (Austermann et al., 2005; Chaavan & Gohokar, 2012; Pao et al., 2007; Yu, 2008) emotions.
- Most of the speech corpora used for emotion recognition are available in languages different from Spanish. Hence, speech

Table 1
Expression-based emotion recognition systems: Ne = Neutral, Fe = Fear, Ha = Happiness, Sa = Sadness, An = Anger, Di = Disgust, Su = Surprise, Bo = Bored.

Work	Emotions	Pattern	Database	Classification technique	Recognition rate
(Karthigayan et al., 2008)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression	Own (1 subject)	ANN	83.57%–85.13%
(Luo et al., 2013)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression	Own (1 subject)	SVM	93.75%
(Filko & Martinovic, 2013)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression	FEEDTUM (18 subjects)	ANN	70.00%
(Pooja & Kaur, 2010)	4 (Ne, Fe, Ha, An)	Facial expression	JAFFE (10 subjects)	ANN	90.00%
(Kaur et al., 2010)	5 (Ha, Sa, An, Di, Su)	Facial expression	JAFFE (10 subjects)	PCA (Euclidean Distance)	80.00%
(Thuseethan & Kuhanesan, 2014)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	JAFFE (10 subjects)	PCA (Euclidean Distance)	91.16%
(Gosavi & Khot, 2013)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	JAFFE (10 subjects)	PCA (Euclidean Distance)	91.19%
(Zhang et al., 2013)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	Cohn-Kanade CK+ (123 subjects)	ANN	78.60%-79.50%
(Zavaschi et al., 2013)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression	JAFFE (10 subjects), Cohn-Kanade CK+ (123 subjects)	SVM	96.20%-99.40%
(Owusu et al., 2014)	(a) 7 (Ne, Fe, Ha, Sa, An, Di, Su), (b) 4 (Ne, Ha, Sa, Su)	Facial expression	(a) JAFFE (10 subjects), (b) Yale (15 subjects)	ANN	(a) 96.83%, (b) 92.22%
(Zhang et al., 2015)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	(a) Bosphorus 3D Batabase (105 subjects: 60 men and 45 women), (b) 11 subjects for live tests	SVM	(a) 92.2%, (b) 84.0%
(Ali et al., 2015)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression	JAFFE (10 subjects), Cohn-Kanade CK+ (123 subjects)	k-NN, SVM, ANN	99.75%
(Rao et al., 2011)	5 (Ne, Fe, Ha, Sa, An)	Facial expression	Own (20 subjects)	ANN	87.00%
(Tayal & Vijay, 2012)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression	FACES collection	PCA (Euclidean Distance)	96.66%
(Chaturvedi & Tripathi, 2014)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	Cohn-Kanade CK+ (123 subjects)	Fuzzy Logic	87.67%
(Pal & Hasan, 2014)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	FEEDTUM (18 subjects)	Fuzzy Logic	89.33%
(Ilbeygi & Hosseini, 2012)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	RaFD (67 subjects)	Fuzzy Logic	93.96%
(Rasoulzadeh, 2012)	6 (Fe, Ha, Sa, An, Di, Su)	Facial expression	JAFFE (10 subjects)	Fuzzy Logic	92.33%

Table 2
Voice-based emotion recognition systems: Ne = Neutral, Fe = Fear, Ha = Happiness, Sa = Sadness, An = Anger, Di = Disgust, Su = Surprise, Bo = Bored.

Work	Emotions	Pattern	Database	Classification technique	Recognition rate
(Shing et al., 2014)	6 (Fe, Ha, Sa, An, Di, Su)	Voice	eINTERFACE'05 (42 subjects), RML (720 samples)	ANN	75.98%–68.57%
(Chen et al., 2012)	6 (Fe, Ha, Sa, An, Di, Su)	Voice	BHUDES (15 subjects, Mandarin Chinese, 20 phrases per emotion)	SVM, ANN	39.16%–50.00%
(Chaavan & Gohokar, 2012)	5 (Ne, Fe, Ha, Sa, An)	Voice	DES (4 subjects, Danish, 88 phrases)	SVM	64.77%–79.55%
(Yu et al., 2001)	4 (Ne, Ha, Sa, An)	Voice	TV recordings (Mandarin Chinese, 721 phrases)	SVM	74.28%
(Austermann et al., 2005)	5 (Ne, Fe, Ha, Sa, An)	Voice	Own (4 subjects, German, 260-280 phrases)	Fuzzy Logic	84.00% (Dependent), 60.00% (Independent)
(Caballero, 2013)	4 (Ne, Ha, Sa, An)	Voice	Own (6 subjects, Spanish, 10 phrases per emotion)	HMM	94.32%
(Schuller et al., 2003)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Voice	Own (5 subjects, German and English, 100 phrases per emotion)	HMM	86.00%
(Pao et al., 2007)	5 (Ne, Bo, Ha, Sa, An)	Voice	Own (34 subjects, Mandarin Chinese, 3400 phrases)	HMM	62.50%
(Yun & Yoo, 2009)	7 (Ne, Bo, Fe, Ha, Sa, An, Di)	Voice	BERLIN (10 subjects, German, 10 phrases per emotion)	HMM	89.00%
(Yu, 2008)	5 (Ne, Fe, Ha, Sa, An)	Voice	Own (Mandarin Chinese)	HMM	87.00%
(Lee et al., 2004)	4 (Ne, Ha, Sa, An)	Voice	Own (1 subject, English, 151-263 phrases per emotion)	HMM	76.12%
(Schuller et al., 2010)	5 - 7	Voice	Corpora (mixture of 6 corpora)	SVM	81.00%
(Vlasenko et al., 2007)	7 (Ne, Bo, Fe, Ha, Sa, An, Di)	Voice	BERLIN (10 subjects, German, 10 phrases per emotion), SUSAS (32 subjects, English)	SVM, HMM	90.00% (BERLIN), 83.00% (SUSAS)
(Wu & Liang, 2011)	4 (Ne, Ha, Sa, An)	Voice	Own (2033 phrases)	GMM, SVM, ANN	72.61% (GMM), 75.33%–78.16% (SVM), 69.86%-71.87% (ANN)
(Firoz-Shah et al., 2009)	4 (Ne, Ha, Sa, An)	Voice	Own (Malabari, 700 phrases)	ANN	55.00%–68.50%

Table 3
Multimodal emotion recognition systems: Ne = Neutral, Fe = Fear, Ha = Happiness, Sa = Sadness, An = Anger, Di = Disgust, Su = Surprise, Bo = Bored.

Work	Emotions	Pattern	Database	Classification technique	Recognition rate
(Busso et al., 2004)	4 (Ne, Ha, An, Sa)	Facial expression-voice	Own (1 actress, 612 phrases)	PCA, SVM	Speech system: 71.00%, Vision system: 85.00%, Multimodal system: 89.00%
(Haq et al., 2008)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression-voice	Own (1 male subject, 120 phrases)	PCA, LDA	Speech system: 53.00%, Vision system: 98.00%, Multimodal system: 98.00%
(Wan & Guan, 2005)	6 (Ne, Fe, Ha, Sa, An, Su)	Facial expression-voice	Own (500 videos, 8 subjects)	LDA	Multimodal system: 82.00%
(Schuller et al., 2007)	3 (Ha, Sa, An)	Facial expression-voice	Own (10 female subjects, 11 male subjects, 10.5 hrs of spontaneous conversation)	SVM	Multimodal system: 64.00%
(Song et al., 2004)	7 (Ne, Fe, Ha, Sa, An, Di, Su)	Facial expression-voice	Own (1384 phrases/samples)	THMM	Multimodal system: 85.00%

corpora are mainly available in East Asian (Chen et al., 2012; Firoz-Shah et al., 2009; Yu et al., 2001; Yu, 2008), Danish (Chaavan & Gohokar, 2012), German (Alter et al., 2000; Schuller et al., 2003) and English (Batliner et al., 2004; Lee et al., 2004; Schuller et al., 2003) languages.

- For training and testing purposes, works on facial expression recognition have considered one subject (Karthigayan et al., 2008; Luo et al., 2013), ten subjects (Gosavi & Khot, 2013; Kaur et al., 2010; Pooja & Kaur, 2010; Rasoulzadeh, 2012; Thuseethan & Kuhanesan, 2014), 20 subjects (Rao et al., 2011) and more than 20 subjects (Chaturvedi & Tripathi, 2014; Ilbeygi & Hosseini, 2012). However, as presented in Table 1 most of the works on the development of vision systems have considered ten subjects. In contrast, most of the works on speech emotion recognition have considered one subject (Lee et al., 2004), four to six subjects (Austermann et al., 2005; Caballero, 2013; Chaavan & Gohokar, 2012; Schuller et al., 2003), ten to 15 subjects (Chen et al., 2012; Vlasenko et al., 2007; Yun & Yoo, 2009) and 34 subjects (Pao et al., 2007). As presented in Table 2 most of the works on the development of speech systems have considered a number of four to ten subjects.
- Works on multimodal systems have considered one subject (Busso et al., 2004; Haq et al., 2008), eight subjects (Wan & Guan, 2005) and 21 subjects (Schuller et al., 2007).

As discussed, most of the works performed in emotion recognition consider some (or all) the following elements to validate their approaches: (1) standard databases or limited volunteers for live tests, (2) a single system (unimodal, or multimodal) which is expected to recognize all considered emotions, (3) a fixed structure for the recognition technique (ANNs, HMMs, etc.), and (4) speech emotion is detected but no sentence is recognized at the same time in order to perform a emotion-dependent conversation with an artificial entity.

These observations represent opportunities for improvement on the topic of emotion recognition research. In this work we (1) explored on the implications of using standard databases for assessment of emotion recognition techniques, (2) extended on the evolutionary optimization of ANNs and HMMs for the development of a multimodal emotion recognition system, (3) set the guidelines for the development of emotional databases of speech and facial expressions, including a set of rules for phonetic transcription of Mexican speech, and (4) evaluated the suitability of the multimodal system within the context of spoken dialogue between a humanoid robot and human users.

Because it is important to improve emotion recognition performance and integrate these recognition systems into robotic entities to achieve an efficient and intelligent communication for human-robot interaction (Kulic & Croft, 2007; Shin et al., 2006). This leads to the contributions of the present work in the field of emotion recognition and human-computer interaction which are highlighted as follows:

- to establish the parameters of an audio-visual database for development of multimodal emotion recognition systems (in this work a database with Mexican users was created);
- to explore on the use of Genetic Algorithms (GAs) to estimate the most suitable structures for ANNs and HMMs for modelling of speech and visual emotional features;
- to analyze the performance of ANNs and HMMs for recognition of speech and visual emotional features with different testing/training schemes;
- to define an approach to integrate a vision and speech emotion recognition system for the development of a multimodal system;
- to establish a guidance for the development of a dialog system for the management of the multimodal recognition system;
- to develop an electronic system to link the multimodal and dialog systems with the humanoid robot Bioloid for human-robot interaction.

Also, the development of intelligent systems for emotion recognition can be improved by the findings of the present work:

- emotion recognition depends on the structure of the database sub-sets used for training and testing, and it also depends on the type of technique used for recognition where a specific emotion can be highly recognized by a specific technique;
- the rules for phonetic transcription of Mexican speech can provide suitable training data for recognition of emotional speech;
- optimization of HMMs led to a Bakis structure which is more suitable for acoustic modeling of emotion-specific vowels;
- some emotions can be better recognized based on speech patterns instead of visual patterns;
- the weighted integration of the multimodal emotion recognition system optimized with these observations can achieve a recognition rate up to 97.00% in live dialogue tests with a humanoid robot.

In Fig. 2 the modules and techniques of the proposed multimodal system are presented. Live tests performed with the multimodal system presented an emotion recognition performance of 97.00% with independent users. In the following sections detailed information of each module is presented and discussed.

2. Audio-visual emotional database: MX database

In order to develop an emotion recognition system a database is needed. As presented in Tables 1 and 2 most of the works on vision systems have used standard databases like JAFFE and FEED-TUM; therefore, few works have built their own database for experiments. For the development of speech systems the available

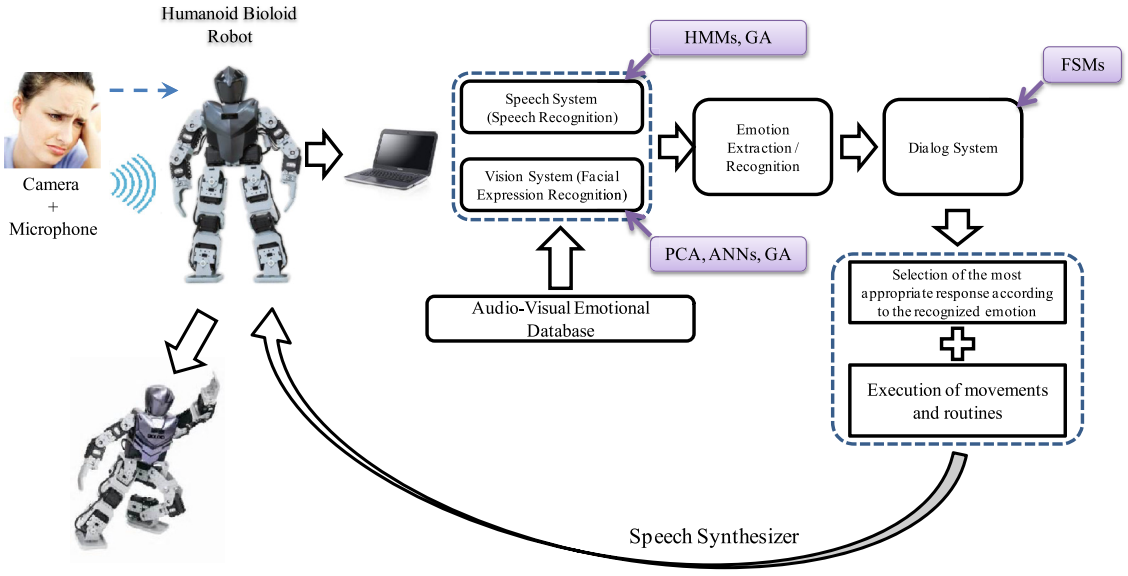


Fig. 2. Modules and techniques of the multimodal emotion recognition system.

databases are presented in other languages (e.g. English, Danish, German, Mandarin Chinese) which are not easily adaptable for the recognition of Mexican Spanish speech due to phonetic and pronunciation differences. The availability of databases with users of different cultural backgrounds is important because depending of these backgrounds there may be differences in emotional facial expressions and voice intonations.

For the development of the present work a database of emotional speech and facial expressions was built with Mexican users. Facial and speech samples were collected from Eastern and Southwestern regions of Mexico to provide a diversity (Pérez-Gaspar et al., 2015a; 2015b). A relevant characteristic for the database was to keep spontaneous emotion performance from non-professional actors to provide near-natural features. In addition, lighting and noise conditions were less strict during collection of facial and speech samples.

In order to keep consistency with previous works the following main emotions were considered (Caballero, 2013; Firoz-Shah et al., 2009; Lee et al., 2004; Wu & Liang, 2011; Yu et al., 2001): Anger (AN), Happiness (HA), Neutral (NE) and Sadness (SA). Because as presented in Tables 1, 2 and 3 most of the databases for emotion recognition consider approximately ten subjects a similar number was considered for the present work.

The Mexican database is separated into two groups: MX-Expressions and MX-Speech. The MX-Expressions database contains pictures from nine subjects (three males and six females) whose details are presented in Fig. 3 (Pérez-Gaspar et al., 2015a). While the samples of the database were in color the samples were pre-processed in gray scale with face extraction to keep consistency with the characteristics of the JAFFE database. Face extraction was performed with the algorithm presented in (Viola & Jones, 2004) and three pictures were captured for each emotion, leading to 12 pictures per subject (108 pictures in total) (Pérez-Gaspar et al., 2015a).

The MX-Speech database contains speech samples from eight subjects (three males and five females) whose details are presented in Fig. 4. For the creation of this database the following conditions were considered (Caballero, 2013; López et al., 2006; Pérez-Gaspar et al., 2015b):

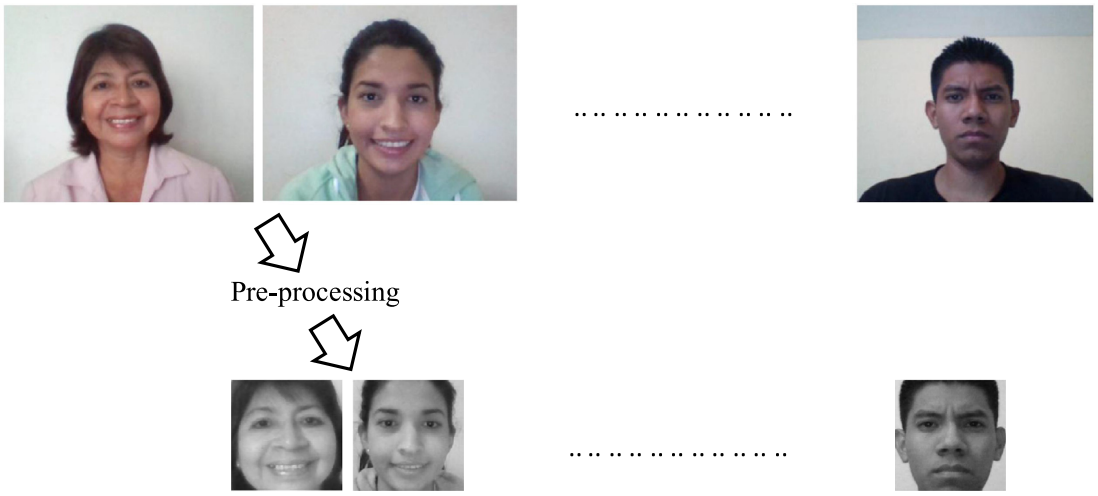
- text stimuli of different length for each emotion;
- semantic significance of the text stimuli;
- there must be enough occurrence of emotion-specific vowels and consonants in the text stimuli.

The text stimuli was designed to help the users to express each emotion. Hence, the stimuli text for Anger, Happiness and Sadness consisted of phrases that were set within the context of daily life situations. For Neutral, the phrases consisted of general information. For this work, 20 stimuli phrases were designed for each emotion, leading to 80 phrases per user (640 phrases in total). All stimuli phrases are presented in Fig. 4.

The speech samples were recorded in .WAV format (at 48000 Hz) and the distance between the microphone (internal laptop microphone) and the user was set to about 60 cm. Then the audio files were labeled at the word and phonetic levels with the software *WaveSurfer* for supervised training of the recognition techniques. Finally the speech samples were coded into Mel Frequency Cepstral Coefficients (MFCCs) since this format enables faster processing with an efficient audio compression (Davis & Mermelstein, 1980; Young & Woodland, 2006). The front-end used 12 MFCCs feature vectors plus energy, delta and acceleration coefficients.

In order to identify the words and (subsequently) the vowel phonemes which were uttered with a particular emotion an identifier was added to the word and phonetic labels. For each emotion the identifier for words was "_E" for Anger, "_F" for Happiness, "_N" for Neutral and "_T" for Sadness.

The phonetic definitions required for the phonetic labeling were obtained from the Mexican Spanish alphabet defined for the transcription of the speech corpus DIMEX100 (Pineda et al., 2010). This alphabet identified 27 phonemes (22 consonants + 5 vowels) for the Mexican Spanish language. Because the emotion recognition approach is based on the evidence that pronunciation of vowels have information regarding the emotions, vowels were considered emotion-dependent. Thus, for phonetic labeling, the following identifiers were added to the vowels of the words spoken with Anger, Happiness, Neutral and Sadness respectively: "_e", "_f", "_n" and "_t" (Caballero, 2013; Pérez-Gaspar et al., 2015b). Because a set of vowels was considered for each emotion a total of 20 vowels (five vowels per four emotions) were integrated into the phonetic repertoire for the Mexican Spanish language leading to 42 phonemes (22 consonants and 20 vowels). In Table 4 the frequency of each emotion-specific vowel in the text stimuli is presented.



User	Au	Me	Je	Ta	Ne	Mi	Fe	Ke	Lu
Gender (Male, Female)	F	F	M	F	F	M	F	F	M
Age (Years)	53	20	21	32	32	30	12	10	25
Region of Origin	Veracruz	Veracruz	Veracruz	Oaxaca	Oaxaca	Oaxaca	Veracruz	Veracruz	Veracruz

Fig. 3. Profile of the participants of the MX-Expressions database.

Table 4
Frequency of emotion-specific vowels per emotional text stimuli.

Vowel	Anger ("_e")	Happiness ("_f")	Neutral ("_n")	Sadness ("_t")
a	65	86	92	83
e	83	94	115	86
i	38	46	60	58
o	54	54	74	65
u	23	28	35	19

An important resource to obtain the phonetic labels for any word, enabling adaptation of large vocabulary, a phonetic transcriptor was developed considering the definitions of the tool TranscribEMex (Pineda et al., 2010). A set of grammatical and acoustic rules were defined to achieve reliable phonetic transcription from the alphabet letters of a word. This was important because the representation of the sound (phoneme) associated to some letters/characters depends on their contexts (e.g., the letters that are inserted before and after). Fig. 5 presents the 50 grammar and acoustic rules defined for the phonetic transcription of different consonant and vowel combinations within a word.

In previous work we reported on the development of the individual vision and speech systems with these databases (Pérez-Gaspar et al., 2015a; 2015b). In the following section an overview of the development of these systems is presented.

3. Individual emotion recognition systems

3.1. Vision system

Fig. 5 presents a general overview of the approaches considered for the development of the Vision System (Pérez-Gaspar et al., 2015a). These approaches were the following:

- Recognizer based on PCA (PCA System): In this approach PCA is applied for dimensionality reduction and identification of the emotion. This is an extension of the eigenface approach (PCA for face recognition) described in (Turk & Pentland, 1991). In the eigenface approach, face identification is performed by computing the Euclidean distance between weights that represent the eigenfaces of a set of pictures that are labeled to identify users, and an input image projected within the eigenface space. The minimum distance identifies the eigenface that better describes the input image and consequently the identification of the user. If the label of the picture consists of only the emotional state expressed by the face (regardless of the user's identity) then the eigenface approach can be used for emotion recognition (Pérez-Gaspar et al., 2015a).
- Recognizer based on PCA and ANN (PCA + ANN System): In this approach PCA is applied for dimensionality reduction which produces the weights that represent the eigenfaces of the MX-Expressions database. These weights are modeled with an ANN to perform the recognition task instead of using the Euclidean distance (Pérez-Gaspar et al., 2015a).
- Recognizer based on PCA, ANNs, and GA (PCA + ANN + GA System): In this approach a GA is applied to identify the most suitable structure of the ANN of the PCA+ANN System.

Fig. 6 presents the details of the GA designed to optimize the structure of the ANN. For the emotion recognition experiments cross-validation was performed. Because there were three picture samples per emotion in the MX-Expressions (numbered as "1", "2", and "3") the cross-validation schemes presented in Table 6 were defined for training, GA optimization, and system evaluation (testing). Table 7 presents the recognition performance for each scheme and recognition system and the following patterns were observed:

- Higher recognition rates for Anger, Happiness and Neutral are obtained with the integrated PCA + ANN + GA System in com-

User	Au	Me	Lu	Je	Li	Ta	Ne	Mi
Gender (Male, Female)	F	F	M	M	F	F	F	M
Age (Years)	53	20	25	21	16	32	32	30
Region of Origin	Veracruz	Veracruz	Veracruz	Veracruz	Veracruz	Oaxaca	Oaxaca	Oaxaca

Speech Recording

Stimuli Text for each User

Anger Phrases

1. Que yo no tuve la culpa.
2. Te digo que te vayas.
3. Mira… haz lo que quieras desde este momento.
4. Yo no te voy a estar soportando.
5. Nunca me incluyes en tus planes.
6. Ya me tienes harto, ya deja de hablar.
7. Ya no te quiero volver a ver.
8. El maestro nos reprobó a todos por tu culpa.
9. Jamás te vuelvas a acercar a mí.
10. Y ahora ¡¿qué es lo que quieres?!
11. Se me olvido mi dinero y tengo que regresarme otra vez.
12. ¡Firulais deja de estarme mordiendo!
13. Te explique muchas veces y no me entiendes.
14. Ya no me pidas dinero.
15. Me cancelaron mi servicio de cable y ya lo había pagado.
16. Esta es una situación que no está a discusión.
17. ¡Corre que la ayuda es urgente inútil !
18. Tú solo me usas a tu conveniencia.
19. Hubieras luchado un poco más, ¡pero no quisiste!
20. No hagas eso que estás de luto.

Neutral Phrases

1. Paris es la capital de Francia.
2. Estudio en la Universidad Tecnológica de la Mixteca.
3. Tengo la edad de 25 años.
4. El año tiene 4 estaciones.
5. La letra "j" es la única que no aparece en la tabla periódica.
6. La Mona Lisa es una obra de Leonardo da Vinci.
7. El océano Atlántico es más salado que el Pacífico.
8. El sistema solar tiene 8 planetas.
9. La araña Sidney es la más venenosa y puede matar un humano en 15 minutos.
10. El graznido de un pato no hace eco.
11. Los mosquitos tienen 47 dientes.
12. El monte Everest es el más alto del mundo.
13. No se puede estornudar con los ojos abiertos.
14. La velocidad del estornudo oscila entre 70 y 150 km/h.
15. El edificio Burj en Dubai es el más alto del mundo con 828 metros.
16. Uruguay es un país de América del Sur.
17. Un avestruz es un ave que no vuela.
18. Una úlcera gástrica ocurre en el estómago.
19. La urticaria es una enfermedad de la piel.
20. El universo es muy grande.

Happiness Phrases

1. ¡Me gane un viaje todo pagado a Florida!
2. Saqué 10 de promedio en mi escuela.
3. ¡Me compre un billete de lotería y gané!
4. Estoy feliz porque me ascendieron de puesto.
5. ¡Tuve unas súper vacaciones!.
6. Que deliciosa estuvo mi comida.
7. Me regalaron pases dobles para ir al cine.
8. Ganamos el primer lugar en el torneo de futbol.
9. Me regalaron un auto en mi cumpleaños.
10. Gané una beca para estudiar en el extranjero.
11. ¡Me encontré 500 pesos tirados en la calle!
12. Ya casi es navidad y veré a toda mi familia junta otra vez.
13. Esta noche iré a cenar con mis amigos y después al cine.
14. Me regalaron el celular que tanto quería.
15. ¡Me regalaron una pantalla 3D extraordinaria!.
16. ¡Es un día fabuloso!
17. ¡Sucedió el milagro que esperaba!
18. Hemos logrado una victoria más.
19. Jugaremos todo el día, es genial.
20. Nadé con una tortuga galápagos.

Sadness Phrases

1. Quisiera volver el tiempo para ver a mi familia.
2. Por más que se hizo no se pudo salvar mi perrito.
3. No entiendo porque ya no quieres verme.
4. Nadie me comprende ni me escucha.
5. Me siento solo, no tengo a nadie con quien platicar.
6. Mi vida es un fracaso… ya no puedo seguir.
7. Por favor dame otra oportunidad.
8. Quisiera que entendieras lo que yo siento.
9. Reprobé todas las materias y me sacarán de la escuela.
10. Mi mejor amigo acaba de fallecer ayer.
11. Acaban de atropellar mi gatito recién nacido.
12. Maté a mi perrito porque no tenía como alimentarlo.
13. Quisiera dormir y no despertar nunca.
14. Me haces mucha falta te extraño.
15. Tu recuerdo sigue aquí todavía.
16. Tuve una sensación como si estuvieras conmigo.
17. La vida es dura y el universo conspira contra mí.
18. Trato de unir los pedazos de mi corazón.
19. Siento que me hundo en un mar de lágrimas.
20. Tu crueldad está acabando con mi mundo.

Fig. 4. Profile of the participants of the MX-Speech database and text stimuli.

Table 5
Rules for the phonetic transcriptor.

Rule	Letter	Context (α) + Letter + (β)	Phonetic representation z = e,f,n,t	Type
1	A	(∗) + A + (∗)	/a_z/	vowel
2	E	(∗) + E + (∗)	/e_z/	vowel
3	I	(∗) + I + (∗)	/i_z/	vowel
4	O	(∗) + O + (∗)	/o_z/	vowel
5	U	(∗)+U+(∗)	/u_z/	vowel
		Specific cases:		
6		(G)+U+(I)	it is not transcribed	
7		(Q)+U+(I)	it is not transcribed	
8		(G)+U+(E)	it is not transcribed	
9		(Q)+U+(E)	it is not transcribed	
10	H	(∗) + H + (∗)	it is not transcribed	
11	B	(∗) + B + (∗)	/b/	consonant
12	V	(∗) + V + (∗)	/b/	consonant
13	S	(∗) + S + (∗)	/s/	consonant
14	Z	(∗) + Z + (∗)	/s/	consonant
15	Q	(∗) + Q + (∗)	/k/	consonant
16	K	(∗) + K + (∗)	/k/	consonant
17	F	(∗) + F + (∗)	/f/	consonant
18	P	(∗) + P + (∗)	/p/	consonant
19	J	(∗) + J + (∗)	/x/	consonant
20	Ñ	(∗) + Ñ + (∗)	/nn/	consonant
21	T	(∗) + T + (∗)	/t/	consonant
		Specific cases:		
22		(∗) + T + (B)	/_D/	consonant
	Y	Specific cases:		
23		Ending of a word	/i_z/	vowel
24		Beginning of a word	/Z/	consonant
25		(consonant)+Y+(consonant)	/i_z/	vowel
26		(vowel)+Y+(vowel)	/Z/	consonant
27	L	(∗) + L + (∗)	/l/	consonant
		Specific cases:		
28		(∗)+L+(L), (L)+L+(∗)	/Z/	consonant
29	D	(∗) + D + (∗)	/_D/	consonant
		Specific cases:		
30		(∗)+D+(vowel), (∗)+D+(R)	/d/	consonant
31	G	(∗)+G+(∗)	/_G/	consonant
		Specific cases:		
32		(∗)+G+(A), (∗)+G+(U), (∗)+G+(O), (∗)+G+(R), (∗)+G+(L)	/g/	consonant
33		(∗)+G+(E), (∗)+G+(I)	/x/	consonant
34	N	(∗)+N+(∗)	/n/	consonant
		Specific cases:		
35		Ending of a word	/_N/	consonant
	R	Specific cases:		
36		Ending of a word	/_R/	consonant
37		Beginning of a word	/r/	consonant
38		(∗)+R+(R), (R)+R+(∗)	/r/	consonant
39		(∗)+R+(vowel)	/r(/	consonant
	C	Specific cases:		
40		(∗)+C+(A), (∗)+C+(O), (∗)+C+(U), (∗)+C+(R), (∗)+C+(L)	/k/	consonant
41		(∗)+C+(E), (∗)+C+(I)	/s/	consonant
42		(∗)+C+(T)	/_G/	consonant
43		(∗)+C+(H)	/tS/	consonant
44	M	(∗)+M+(∗)	/m/	consonant
		Specific cases:		
45		(∗)+M+(P)	/_N/	consonant
	X	Specific cases:		
46		Beginning of a word	/x/	consonant
47		(∗)+X+(T), (∗)+X+(P), (∗)+X+(C)	/kS/	consonant
48		(vowel)+X+(vowel), Word begins with a vowel	/x/	consonant
49		(vowel)+X+(vowel), Word begins with a consonant	/kS/	consonant
50	W	(∗)+W+(∗)	/g/ + /u_z/	consonant + vowel

α = letter inserted before the letter to be transcribed, β = letter inserted after the letter to be transcribed, ∗ = any letter (vowel or consonant).

Table 6
Cross-validation schemes to determine the best approach for the development of the vision system.

Scheme	S_1	S_2	S_3	S_4	S_5	S_6
Training	1	1	2	2	3	3
GA optimization	3	2	3	1	2	1
Testing	2	3	1	3	1	2

parison with the PCA and PCA + ANN Systems. For Sadness higher recognition rate is obtained with the PCA System.

- For the PCA + ANN + GA System an ANN structure was obtained for each cross-validation scheme. The total average recognition rate obtained with these specific ANN structures was 79.36%. By estimating an average of these structures the mean structure "[1 57 3]" (i.e., one hidden layer with 57 neurons per layer

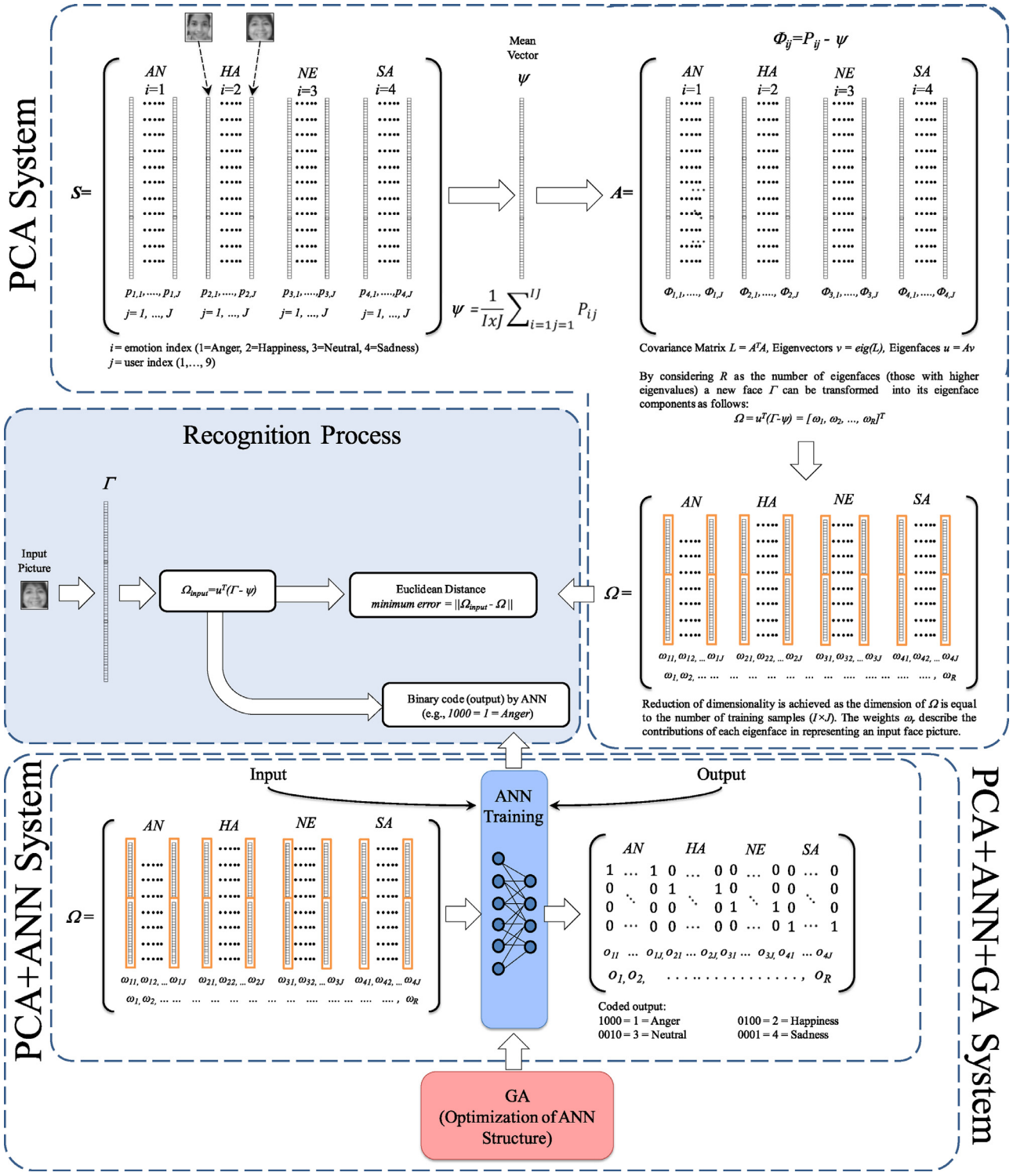


Fig. 5. Overview of the approaches for the development of the vision system.

and linear transfer function) was obtained. This mean structure increased the recognition rate for all cross-validation schemes and emotions, leading to a total average recognition rate of 82.41%.

- Independently of the cross-validation scheme the PCA + ANN + GA System presents a higher recognition rate than the other systems (with specific and mean ANN structures). However there is a significant variability between the recognition rates of these schemes.

Based on these results the PCA + ANN + GA System was considered as the Vision System for multimodal emotion recognition. In the following section an overview of the speech system is presented.

3.2. Speech system

When a speech recognition system is developed, every speech sound must be modeled according to the phonemes involved in

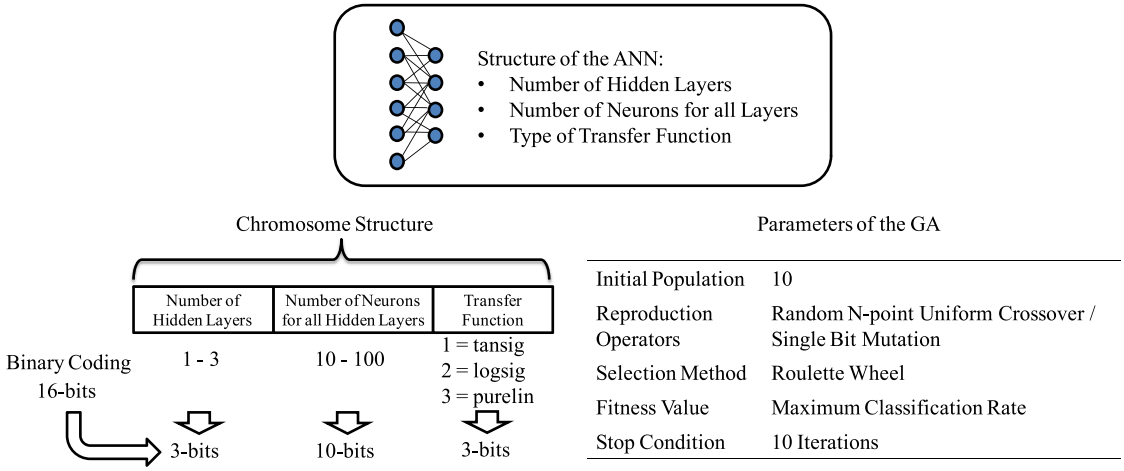


Fig. 6. Chromosome structure and configuration parameters of the Genetic Algorithm for the PCA+ANN system (vision system).

Table 7
Emotion recognition performance of the vision systems on the MX-Expressions database.

System	Training	Optimization	Testing	Estimated ANN Structure	AN	HA	NE	SA	Average[2]
PCA	1		2		77.78	77.78	55.56	77.78	72.23
	1		3		77.78	88.89	55.56	100.00	80.56
	2		1		66.67	88.89	55.56	77.78	72.23
	2		3		55.56	66.67	66.67	100.00	72.23
	3		1		66.67	66.67	55.56	77.78	66.67
	3		2		66.67	77.78	66.67	77.78	72.23
				Average[1] PCA	68.52	77.78	59.26	85.19	72.69
PCA + ANN	1		2		77.78	88.89	77.78	66.67	77.78
	1		3		66.67	88.89	33.33	66.67	63.89
	2		1		55.56	77.78	55.56	100.00	72.23
	2		3		88.89	77.78	66.67	55.56	72.23
	3		1		66.67	77.78	44.44	44.44	58.33
	3		2		88.89	77.78	55.56	77.78	75.00
				Average[1] PCA+ANN	74.08	81.48	55.56	68.52	69.91
PCA+ANN+ GA	1	2	3	[1 71 3]	80.37	88.88	45.55	61.85	69.16
	1	3	2	[2 78 3]	84.07	88.51	72.96	77.40	80.74
	2	1	3	[1 56 3]	88.88	100.00	77.77	79.25	86.48
	2	3	1	[2 64 3]	72.96	97.03	77.77	72.96	80.18
	3	1	2	[1 30 3]	82.59	88.88	77.77	100.00	87.31
	3	2	1	[1 42 3]	59.25	90.00	60.37	79.62	72.31
				a) Average[1] PCA+ANN+AG	78.02	92.22	68.70	78.51	79.36
	1	2	3	Mean Structure= [1.33 56.83 3] ≈ [1 57 3]	88.89	88.89	44.44	66.67	72.22
	1	3	2		88.89	88.89	77.78	77.78	83.34
	2	1	3		88.89	100.00	77.78	88.89	88.89
	2	3	1		77.78	100.00	77.78	88.89	86.11
	3	1	2		88.89	88.89	77.78	100.00	88.89
	3	2	1		55.56	88.89	77.78	77.78	75.00
				b) Average[1] PCA+ANN+AG	81.48	92.59	72.22	83.34	82.41

Average[1] = Average recognition rate per emotion through all cross-validation schemes.
Average[2] = Average recognition rate per cross-validation scheme through all emotions.
Total average recognition rate per system.

the language. Research on emotion recognition (Lee et al., 2004; Li et al., 2010) has demonstrated the effective use of spectral properties of vowel sounds when recognizing an emotion. Intonation of vowels is representative of emotional states and thus a vowel "e" uttered with Anger would have spectral properties different from those of the same vowel uttered with Sadness or Happiness (Caballero, 2013; Pérez-Gaspar et al., 2015b). For this reason, the modeling of emotion-specific vowels (Caballero, 2013; Pérez-Gaspar et al., 2015b) was considered for emotion recognition from speech. As presented in Section 2 the phonetic transcription of the MX-Speech database already considers the identification of emotion-specific vowels.

Fig. 7 presents the standard elements of the speech recognizer that were built with the tool HTK (Young & Woodland, 2006). The acoustic modeling of phonemes was performed with Hidden Markov Models (HMMs) and an HMM was built for each phoneme in the Mexican Spanish language. In addition, HMMs for silence (phoneme /sil/) and short pause (phoneme /sp/) were

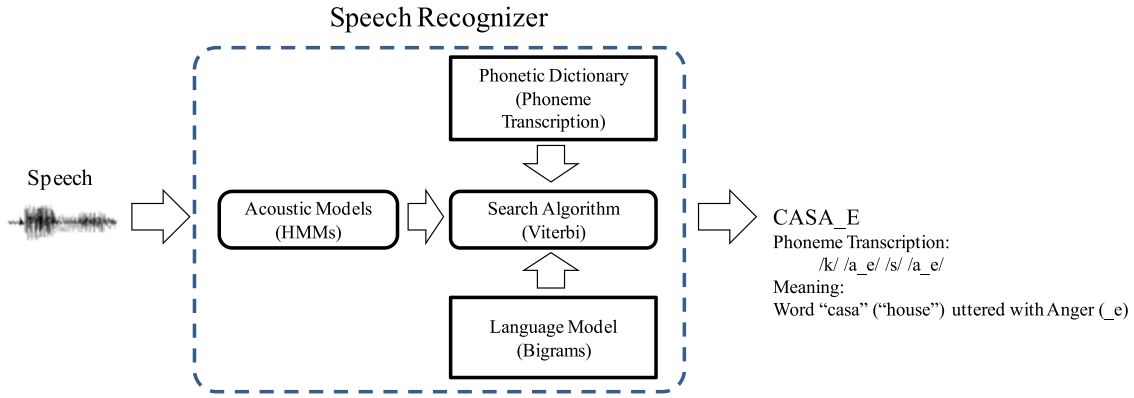


Fig. 7. Standard elements of a speech recognizer.

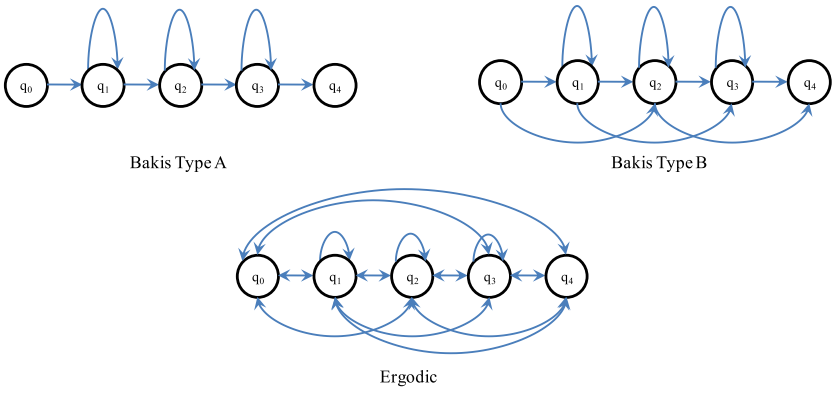


Fig. 8. HMM structures for the acoustic modelling of the emotion-specific vowels.

built. Because the recognizer is based on phonemes a lexicon (phonetic dictionary) is required to restrict the recognized sequence of phonemes to form valid words. These sequences of valid words are then restricted by the language model to form valid sequences of words according to statistical rules. Because any word can be uttered with any emotion the rules of the language model apply to all emotions. Thus, for the experiments performed with the MX-Speech database the language model integrated by the whole set of 80 phrases considering that each phrase could be uttered with all emotions. This led to a total of 80 × 4 emotions = 320 phrases for the estimation of the language model. This was also required to avoid biased recognition of the emotional state given by the language model.

The recognition process that integrates all the elements presented in Fig. 7 is managed by the search algorithm to provide the most suitable word output (transcription) for an acoustic signal of input speech. Emotion recognition on the word output is estimated by counting the number of vowels within the recognized words. The identifier (*_e*, *_f*, *_n*, *_t*) with the higher number of vowels defines the dominant emotion.

In Fig. 8 a set of HMM structures considered for acoustic modeling of phonemes is presented. While the Bakis Type A structure is commonly used (Young & Woodland, 2006) other structures as Bakis Type B or Ergodic may be more suitable for acoustic modeling of emotion-specific vowels. The problem of identifying the appropriate HMM structure for each emotion-specific vowel can be addressed via a GA. Hence, for the development of the speech system two approaches were considered for the acoustic modeling of phonemes (Pérez-Gaspar et al., 2015b):

- Standard HMMs: all HMMs are considered with the standard Bakis Type A structure.
- GA+HMMs: all HMMs for consonants are considered with the standard Bakis Type A structure. A GA is implemented to assign the most suitable HMM structure for each emotion-specific vowel (hence, all HMMs for vowels can be considered with Bakis Type A, Bakis Type B, or Ergodic structures). Fig. 9 presents the details of the GA designed to optimize the structures of the HMMs. For GA optimization sub-sets of phrases from each emotional state were considered for HMM training and GA optimization. As presented in Fig. 4 each subject in the MX-Speech database has 20 phrases per emotional state. This led to defining the following sub-sets: (1) training set (phrases 13 to 20), and (2) optimization set (phrases 7 to 12). Fig. 9 presents the HMM structures estimated for each emotion-specific vowel.

For the final assessment of the systems based on Standard HMMs and GA+HMMs two recognition schemes were considered:

- Test Scheme A (speaker-dependent): under this scheme 40 sentences (10 first sentences × 4 emotions) from each speaker were considered for training of the HMMs in addition to the 560 sentences (20 sentences × 4 emotions × 7 remaining speakers) from the other speakers. Then recognition performance was evaluated on the speaker's remaining 40 sentences (10 last sentences × 4 emotions).
- Test Scheme B (speaker-independent): under this scheme 40 sentences (10 first sentences × 4 emotions) from each speaker were considered for speaker adaptation with Maximum Likelihood Linear Regression (MLLR) (Young & Woodland, 2006). The HMMs were trained only with the 560 sentences (20 sentences × 4 emotions × 7 remaining speakers) from the other speakers. Recognition performance was evaluated on the speaker's remaining 40 sentences (10 last sentences × 4 emotions).

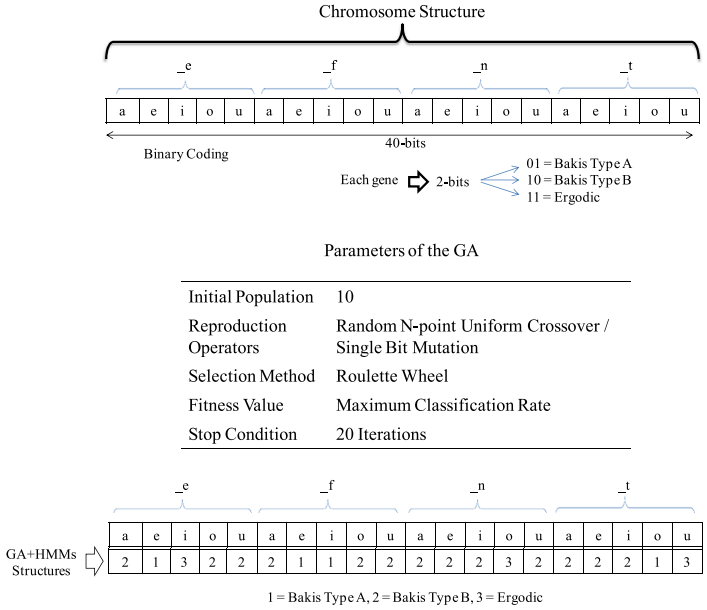


Fig. 9. Chromosome structure, configuration parameters of the Genetic Algorithm for the GA+HMMs System and GA+HMMs structures (speech system).

Table 8
Emotion recognition performance of the speech systems on the MX-Speech database.

System	User	Test Scheme A AN	HA	NE	SA	Average[2]	Test Scheme B AN	HA	NE	SA	Average[2]
Standard HMMs	Lu	100.00	50.00	100.00	80.00	82.50	100.00	50.00	100.00	100.00	87.50
	Ta	100.00	80.00	100.00	90.00	92.50	100.00	70.00	100.00	100.00	92.50
	Au	80.00	85.00	80.00	100.00	86.25	100.00	100.00	80.00	100.00	95.00
	Mi	70.00	70.00	100.00	85.00	81.25	70.00	80.00	100.00	90.00	85.00
	Me	75.00	70.00	90.00	90.00	81.25	95.00	90.00	100.00	100.00	96.25
	Je	100.00	30.00	75.00	50.00	63.75	80.00	100.00	70.00	90.00	85.00
	Li	70.00	40.00	20.00	75.00	51.25	75.00	80.00	75.00	70.00	75.00
	Ne	80.00	100.00	90.00	90.00	90.00	90.00	100.00	100.00	80.00	92.50
	Average[1] StdHMMs	84.38	65.63	81.88	82.50	78.59	88.75	83.75	90.63	91.25	88.59
GA+HMMs	Lu	100.00	60.00	100.00	90.00	87.50	100.00	60.00	100.00	100.00	90.00
	Ta	100.00	90.00	100.00	90.00	95.00	100.00	90.00	100.00	90.00	95.00
	Au	80.00	70.00	80.00	100.00	82.50	100.00	100.00	80.00	100.00	95.00
	Mi	100.00	65.00	100.00	90.00	88.75	70.00	60.00	90.00	90.00	77.50
	Me	65.00	90.00	100.00	90.00	86.25	95.00	100.00	90.00	100.00	96.25
	Je	100.00	20.00	85.00	25.00	57.50	90.00	100.00	90.00	80.00	90.00
	Li	60.00	45.00	80.00	90.00	68.75	90.00	60.00	90.00	70.00	77.50
	Ne	80.00	100.00	100.00	90.00	92.50	100.00	100.00	100.00	80.00	95.00
	Average[1] GA+HMMs	85.63	67.50	93.13	83.13	82.34	93.13	83.75	92.50	88.75	89.53
							Mixed Standard+GA HMMs				
							93.13	83.75	92.50	91.25	90.16

Average[1] = Average recognition rate per emotion through all users.
Average[2] = Average recognition rate per user through all emotions.
Total average recognition rate per system.

In Table 8 the recognition performance of the Standard HMMs and GA+HMMs under both recognition schemes is presented. It is important to mention that for both Standard HMMs and GA + HMMs the speaker-independent scheme was more accurate than the speaker-dependent scheme. Within the speaker-independent scheme the application of the GA improved the recognition performance for Anger and Neutral (Happiness remained unchanged). However for Sadness a decrease in recognition accuracy was obtained. If the Standard HMMs are just considered for the emotion-specific vowels of Sadness the overall recognition rate of the GA+HMMs increases from 89.53% to 90.16% with the speaker-independent scheme. If the overall recognition performance of the Standard HMMs is considered as reference (88.59%) the performance of 90.16% of the Mixed Standard+GA HMMs (i.e., GA + HMMs for Anger, Happiness and Neutral, and Standard HMMs for Sadness) is statistically significant under a non-parametric Wilcoxon test (Pérez-Gaspar et al., 2015b).

The results presented in Table 8 provided support to the use of emotion-specific vowels and the speaker-independent scheme for the development of the speech system. The following section presents the approach to integrate the vision and speech systems for multimodal emotion recognition.

4. Integration of the multimodal emotion recognizer

Multimodal recognition is performed with the integration of the vision and speech systems. This integration provides a global answer to the recognition task considering the strengths of

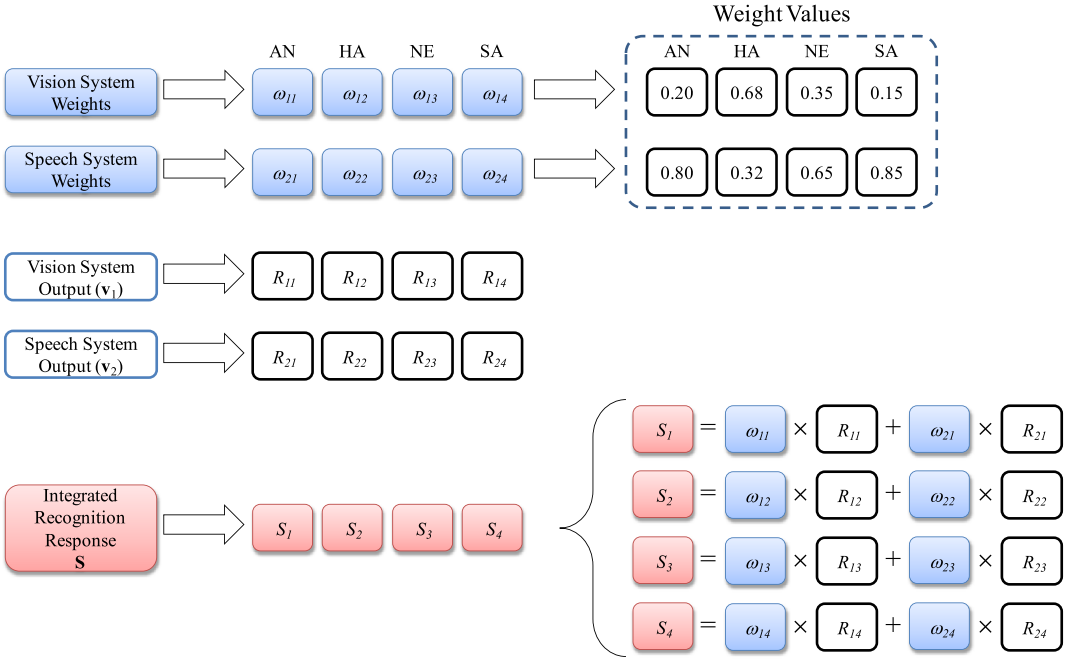


Fig. 10. Weight scheme for the integration of the visual and speech recognition systems for multimodal emotion recognition.

each emotion recognition system. From the results presented in Tables 7 and 8 it was determined that the best configurations for the individual recognition systems were the following:

- Speech system: Standard + GA HMMs with speaker-independent recognition scheme.
- Vision system: PCA + ANN + GA where the ANN had one hidden layer containing 57 neurons and linear activation function.

The response of each individual system can be represented as a row vector with four elements where each element is the recognition performance for a specific emotion. Hence, for a speech/facial expression input, the vector $\mathbf{v} = [60.00, 20.00, 10.00, 10.00]$ would represent the associated output of the recognition system (recognition percentage). This vector shows that the input pattern is identified as: Anger with a probability of 60.00%, Happiness with a probability of 20.00%, Neutral with a probability of 10.00%, and Sadness with a probability of 10.00% (note that 60.00% + 20.00% + 10.00% + 10.00% = 100%). By considering the highest probability or recognition percentage, Anger would be chosen as the correct emotion. However this conclusion could be wrong since the individual recognition system may be prone to confuse a particular emotion with another.

From Tables 7 and 8 it is observed that the speech system performs better for recognition of the emotions of Anger, Neutral and Sadness while the vision system identifies Happiness with higher accuracy. By having this knowledge about the individual systems' performance (e.g., vector $\mathbf{v}$), the integration was performed by means of a weighted sum of recognition responses. Fig. 10 presents the weight model for the integration of recognition responses. Weights are defined as ω_{ij} where i= recognition system index (1 = vision system, 2 = speech system), j is the emotion index (1 = Anger, 2 = Happiness, 3 = Neutral, and 4 = Sadness), and $\sum_{i=1}^{2} \omega_{ij} = 1.0$. The results presented in Tables 7 and 8 were considered to adjust the weights ω_{ij} as presented in Fig. 10.

If $\mathbf{v_1} = [R_{11}, R_{12}, R_{13}, R_{14}]$ and $\mathbf{v_2} = [R_{21}, R_{22}, R_{23}, R_{24}]$ are defined as the output recognition vectors of the vision and speech systems respectively, and integrated recognition response vector ($\mathbf{S}$) can be estimated as:

$$\mathbf{S} = [\omega_{11} \times R_{11} + \omega_{21} \times R_{21}, \omega_{12} \times R_{12} + \omega_{22} \times R_{22}, \omega_{13} \times R_{13} + \omega_{23} \times R_{23}, \omega_{14} \times R_{14} + \omega_{24} \times R_{24}] \quad (1)$$

$$\mathbf{S} = [S_1, S_2, S_3, S_4] \quad (2)$$

It is expected that the weighted sum presented in $\mathbf{S}$ will provide a better insight about the correct emotion, providing the highest S_j to the most likely correct emotion.

4.1. Creation of the dialogue system

The multimodal system generates two outputs: (1) the transcription of the spoken sentence, and (2) the emotion detected by the integration of the speech and vision systems. This information can support emotion-dependent human-robot interaction.

For this purpose a dialogue system was created. Particularly, the dialogue system manages the multimodal emotion recognition system to perform the following tasks:

- to identify the meaning of a sentence spoken by the human user;
- to associate an appropriate response to the identified meaning considering the emotion detected by the multimodal system;
- to execute the appropriate response by means of speech synthesis and robot actions;
- to provide visual and acoustic feedback (robot actions and speech synthesis) to continue the interaction process.

To accomplish these tasks the creation of the dialogue system consisted of the following steps:

- *Definition of a Conversation Context*. This was important to define the vocabulary and sentences for the estimation of the system's language model. For this work, the context identified as "A Day at School" was considered to cover conversations regarding the experiences of a university student during a day at school (e.g., worries about exams' results, complaining about the amount of homework, etc.).

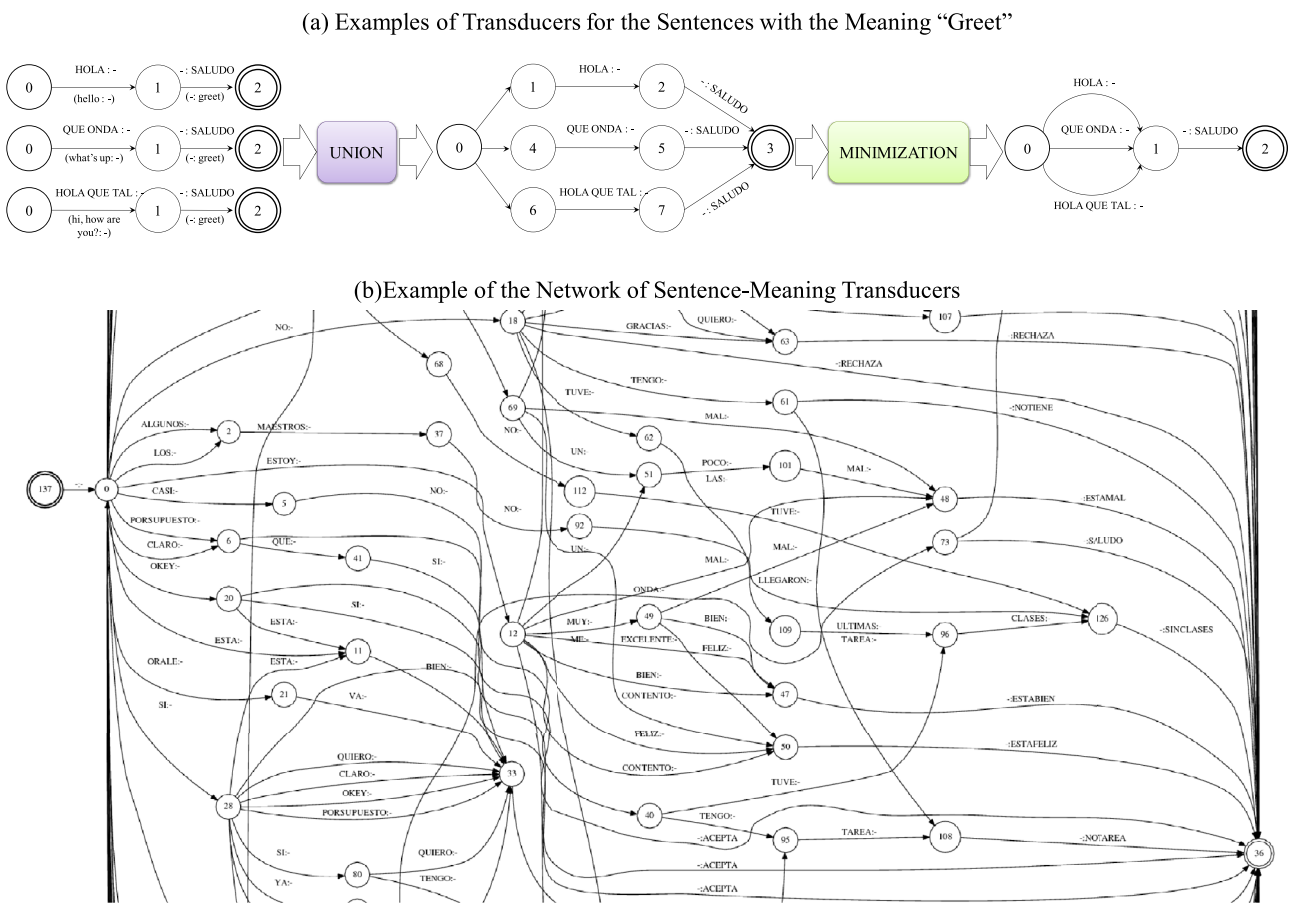


Fig. 11. Examples of (a) sentence-meaning transducers integrated by the union and minimization operators for FSTs, and (b) fragment of the whole Sentence-Meaning FST network.

- *Identify the Vocabulary of the Conversation Context*. In this step all possible sentences that could be present in a conversation regarding the context "A Day at School" were identified. These sentences were obtained from a set of interviews performed to real university students.
- *Define the Sentence-Meaning Relationship for each Sentence*. In this step the set of meanings for each sentence in the conversation context was defined (e.g. the sentence "Hola, ¿cómo estás? = "Hello, how are you?" has the meaning "greet" that requests a "response"). Then, a set of responses were defined according to the emotion detected in the recognized sentence. These responses could consist of (1) only spoken comments via speech synthesis (e.g., "I am fine, thank you for asking, I am glad to see you happy"), (2) spoken comments with additional information requests (e.g., "I am fine, thank you for asking, I noticed you are a little serious, how are you?"), and (3) spoken comments with execution of physical actions by the robot (e.g., "I am fine, thank you for asking, hum, I see you a little sad, let me show you some dance moves").
- *Implementation of Sentence-Meaning Transducers*. Each emotional sentence-meaning relationship was represented with Finite-State Transducers (FSTs), where an input symbol is transformed in terms of a set of output symbols (Mohri, Pereira, F., & Riley, 1996). The implementation through FSTs was performed with the FSM toolkit (AT&T Technology Solutions, 2015). Fig. 11 presents an example of the integration and minimization processes (Mohri et al., 1996) that were performed with the FSM toolkit to integrate all sentence-meaning relationships within a single network of FSTs.

As described, the dialogue system first identifies the spoken sentence and the user's emotional state. Second, a response is proposed according to the meaning and emotional context of the recognized sentence. While spoken responses are provided by the means of speech synthesis, physical responses are performed by the humanoid Bioloid. Some movements and routines designed for the Bioloid robot are described below:

- *Introducing the Robot:* When a new user is detected in front of the camera, the robot greets and says its purpose: to have a conversation to release stress due to a day at school. Movements of the arms emphasizing the greeting are performed.
- *Expressions:*
  - Interrogative Pose: The robot rises its arms with palms up expressing "Why?".
  - Scared: The robot covers its head with its arms.
  - Claps: The robot rises its arms and performs two claps above its head.
  - Affirmative Gestures: The robot up and down its to emphasize an affirmative answer.
  - Joke: The robot says jokes with movements emphasizing the story.
- *Exercise Routines:*
  - Stretching
  - Push-ups
  - Dancing

These routines were developed with the RoboPlus Motion software that comes included with the Bioloid robot (Robotis, 2012).

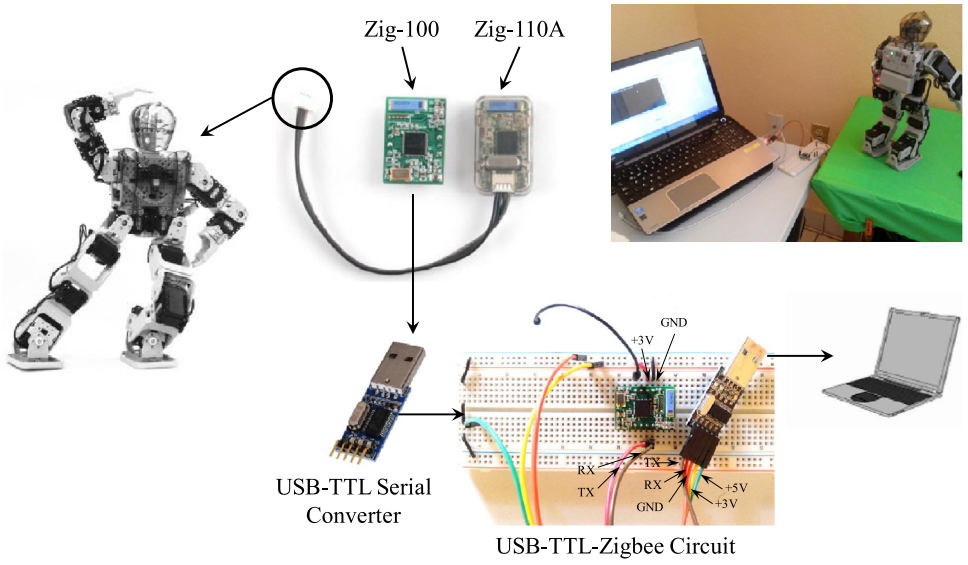


Fig. 12. USB-Zigbee connection for the Bioloid robot.

4.2. Connection of the robot with the multimodal and dialogue systems

As presented, in this work the Bioloid Premium robot in its humanoid version was used (Robotis, 2012). This robot has been widely used by researchers in the field of robotics and it is one of the most used platforms for competition in the Robotics Mexican Tournament and RoboCup (Röfer et al., 2008). It is composed by an Arm Cortex 32-bits CPU and 18 servomotors (dynamixels).

The Bioloid's communication system is based on the Zigbee module to transmit data between the computer and the robot via a wireless connection. Although to establish a wireless connection a USB2Dynamixel and ZIG2Serial modules are commonly used, in this work a USB-TTL converter was used to replace both modules. This led to a more practical and economic connection (Prolific Technology, 2015). The SDK from Robotis was used to establish the connection, modifying a c++ program to receive commands as arguments and to send them to the robot to execute a specific routine. In Fig. 12 the connection with these modules is presented.

5. Graphical user interface and performance results

For live tests, a Graphical User Interface (GUI) that integrated the multimodal system with the dialogue system was developed. This interface incorporates functions for the different tasks required to adapt, train, and use the multimodal system for human-robot interaction. In Fig. 13 the main features of the GUI are presented:

1. *User Recognition*: This button has the purpose to identify the user in front of the camera.
2. *User's Face*: This section shows the extracted face from the main display.
3. *Image Correction*: In this section an equalization can be applied on the user's image. Contrast and brightness can also be adjusted, either automatically (by applying a threshold) or manually.
4. *Recognized Speech*: In this box, the recognized words from the speech system are shown.
5. *Globally Detected Emotion*: In this box, the output of the multimodal system as described in Fig. 10 (see **S**, Eqs. 1 and 2) is presented.
6. *Recognize Emotion*: With this button, the audio and video input devices are activated to start recording speech and visual emotional data. Then, this data is processed to detect the emotion.
7. *Capture of New User's Data*: In this section different buttons are shown. The "C" button enables the function to store the new user's name. Buttons "E","F","N" and "T" have the purpose to capture facial expression samples for Anger, Happiness, Neutral and Sadness respectively. The "Rec" button only applies to the Speech sub-system because it opens a new window where the user can record emotional speech samples for speaker adaptation. Once that all audio-visual samples are captured, the GUI performs visual and speech adaptation of the recognition engines (e.g., ANNs, HMMs, PCA). Speech adaptation is performed following the Test Scheme B described in Section . Visual adaptation is performed with only one sample of facial expression.
8. *Results of the Vision System*: In this box the emotion recognition results of the vision system as described in Fig. 10 (see $\mathbf{v}_1$) are presented.
9. *Results of the Speech System*: In this box the emotion recognition results of the speech system as described in Fig. 10 (see $\mathbf{v}_2$) are presented.
10. *Messages (Process Status)*: In this section, alert or warning messages are shown (e.g., indicates when variables are loaded, or when a sub-system has already been trained).

5.1. Multimodal emotion recognition: live test

For the final validation of the multimodal system, live tests were performed with ten new Mexican users and the GUI. Prior to the live tests for validation, samples of emotional speech and facial expression were collected with the GUI. It is important to mention that these samples were collected under non-controlled conditions to represent a more natural situation when an interaction is performed. The details of the new users (five males and five females) are presented in Table 9.

In Table 10 the results of the live tests are presented. In total, 400 sentences were evaluated in the test sessions (10 users $\times$ 10 sentences $\times$ 4 emotions). As presented, an overall emotion recognition rate of 97.00% was obtained with the multimodal system.

5.2. Human-robot interaction tests

In this section some examples of the interaction routines performed with the human user and the Bioloid robot are presented.

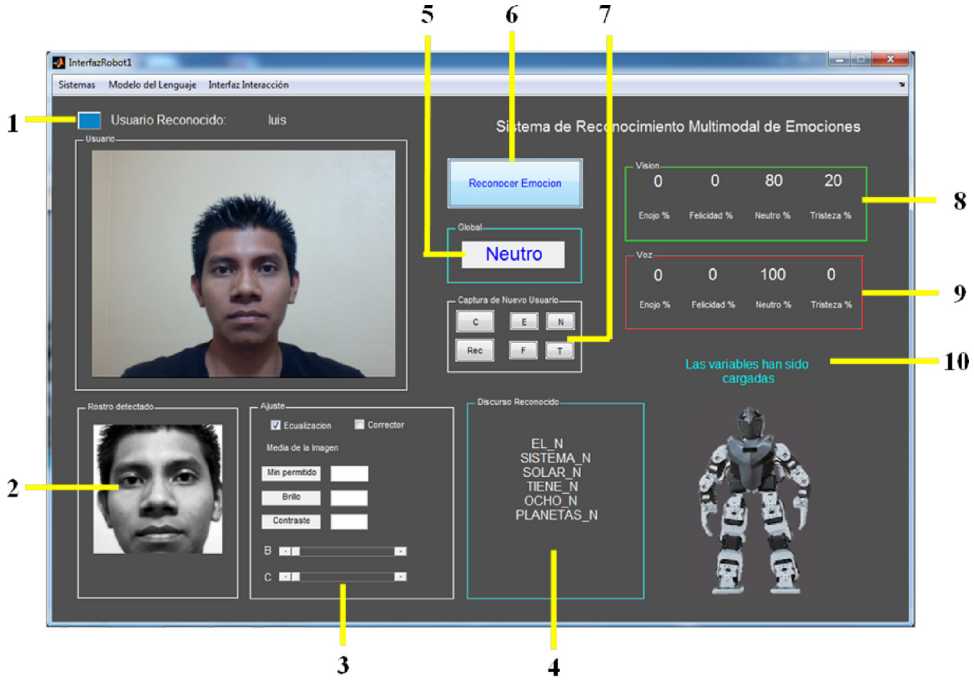


Fig. 13. Capture from the multimodal emotion recognition interface.

Table 9
Profile of the participants for the live tests of the multimodal emotion recognition system.

User	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
Gender (male, female)	M	M	F	M	M	M	F	F	F	F
Age (years)	27	40	22	26	28	27	44	40	45	28
Region of origin	Oaxaca	Oaxaca	Oaxaca	Oaxaca	Oaxaca	Oaxaca	Veracruz	Veracruz	Veracruz	Veracruz

Table 10
Global emotion recognition rate of the multimodal system (live tests).

	AN sentences	HA sentences	NE sentences	SA sentences	Multimodal recognition performance
Users	1 2 3 4 5 6 7 8 9 10	1 2 3 4 5 6 7 8 9 10	1 2 3 4 5 6 7 8 9 10	1 2 3 4 5 6 7 8 9 10	
U1	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	40/40 = 100.0
U2	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ × ✓ ✓ ✓ ✓ ✓ ✓ ✓ ×	38/40 = 95.0
U3	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ × ✓ ✓	✓ × ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	38/40 = 95.0
U4	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	40/40 = 100.0
U5	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	40/40 = 100.0
U6	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ × ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	39/40 = 97.5
U7	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	× ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	39/40 = 97.5
U8	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ × ✓ ✓ ✓ ✓ ✓ × ×	37/40 = 92.5
U9	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	40/40 = 100.0
U10	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ × ✓ × ✓ ✓ ×	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	37/40 = 92.5
	✓ Correct emotion detection	× Incorrect emotion detection		Global emotion recognition rate = 97.0	

- Routine 1:
  - Robot: “Hi, how are you?”
  - User: “I’m upset” (anger tone and expression)
  - Robot: “Don’t be upset, look, check this rotine to release some stress. We start by ... (description of the exercise)...” + (execution of excercise movements). The execution of this routine is presented in Fig. 14.
- Routine 2:
  - Robot: “And tell me, how was school?”
  - User: “I had a bad day” (angry tone and expression)
  - Robot: “Maybe some push-ups can help to improve the bad mood, check how I do them. With strength ... (description of the exercise)...” + (execution of push-up movements). The execution of this routine is presented in Fig. 15.
- Routine 3:
  - Robot: “I want to know something, do you have exams?”
  - User: “Yes, I have” (sad tone and expression)
  - Robot: “I recommend you to start studying so you can get a good grade. Before that, you can hear this joke ... (description of the joke)..., I hope you liked it” + (execution of emphasizing movements). The execution of this routine is presented in Fig. 16.
- Routine 4:
  - Robot: “Tell me, were you given a lot of homework? or not so much?”
  - User: “Yes, I was given too much homework” (sad tone and expression)

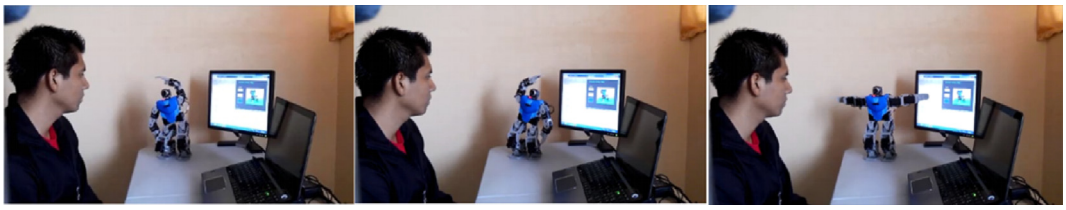


Fig. 14. Dialogue and execution of Routine 1.

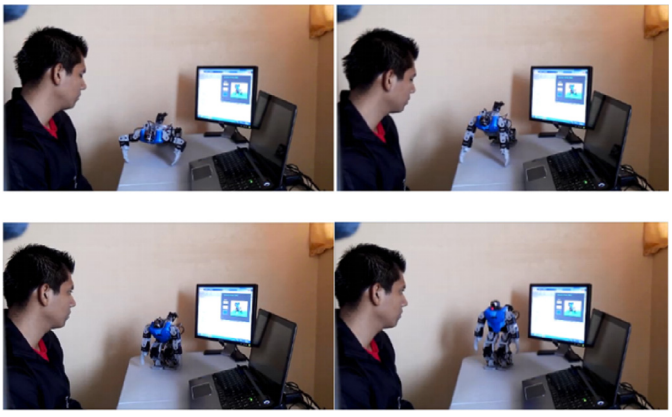


Fig. 15. Dialogue and execution of Routine 2.

Fig. 16. Dialogue and execution of Routine 3.

  - Robot: "Well, start doing your homework and you will see that you will end soon. I can relax you with this dance" + (execution of dance movements) + "I hope you liked this dance". The execution of this routine is presented in Fig. 17.
- Routine 5:
  - Robot: "Hey, you left school early today?"
  - User: "Yes, I left early" (neutral tone and expression)
  - Robot: "That was great. I can entertain you with a joke so you can continue having a nice day ... (description of the joke) ... " + (execution of emphasizing movements). The execution of this routine is presented in Fig. 18.

To verify the consistency of the dialogue system the users expressed with different emotions previously performed routines. For example, routines 2 and 5 were performed by the test user with a different emotion, obtaining the following responses:

- Routine 2:
  - Robot: "And tell me, how was school?"
  - User: "It was very good" (happy tone and expression)
  - Robot: "Look how I cheer at you, everybody must be like this, always keeping a smile in our face" + (execution of clap movements). The execution of this routine is presented in Fig. 19.
- Routine 5:
  - Robot: "Hey, you left school early today?"
  - User: "Yes, I left early" (happy tone and expression)
  - Robot: "Hey, that was good, you had time to come home early and go out with your friends. I am happy as you are" + (execution of King-Kong movements). The execution of this routine is presented in Fig. 20.

In all interaction experiments with different test users the multimodal and dialogue systems provided consistent results.

6. Conclusions and future work

For the integration of the multimodal emotion recognition system the following sub-systems were considered: (1) speech system based on "Standard + GA HMMs", and (2) vision system based on "PCA + ANN + GA". These sub-systems presented the highest recognition rates with a database created with Mexican users (MX-Speech and MX-Expressions). The results obtained during the development of these sub-systems showed that emotion recognition may be dependent of the characteristics of the database, the characteristics of the recognition technique, and the training-testing scheme.

Regarding the development of the speech system, the speaker-independent scheme was determined to present higher emotion recognition rates than the speaker-dependent scheme. Also, the speaker-independent scheme, that requires an adaptation step, was faster to perform than speaker-dependent training. An important observation regarding the performance of the speech system is that recognition rates were consistent through men and women, even though there were more women in the MX-Speech database used for HMM training. This leads to emphasize the good modeling of emotions regardless of gender by focusing on acoustic modeling of emotion-specific vowels and MLLR speaker adaptation. The performance of the speech system was statistically improved by using the evolutionary computation technique of Genetic Algorithms (GA). It was found that the HMM architecture has an important effect on the acoustic modeling of emotion-specific vowels. The structures found by the GA consisted of a combination of Bakis Type A, Type B and Ergodic structures, where the Bakis Type B structure had more presence. However, Bakis Type A was found to be more efficient for modeling the emotion-specific vowels for "Sadness.

The use of GA to optimize the structure of ANNs also led to statistical improvements in the performance of the vision system and the MX-Expressions database. The vision system based on "PCA + ANN + GA" presented the highest emotion recognition rates in facial expressions through different training-testing schemes. However, it was found that the vision system based on PCA could be more efficient for a particular emotion. Hence, a specific technique could be more appropriate for some emotions instead that other techniques. As presented in Table 7 the system based on PCA can be more suitable to recognize "Sadness" and the system based on "PCA + ANN + GA" can be more suitable to recognize "Anger", "Happiness" and "Neutral".

For the purpose of this work, the multimodal system presented a global emotion recognition rate of 97.0% in live tests with Mexican users that were different from those of the MX-Speech and

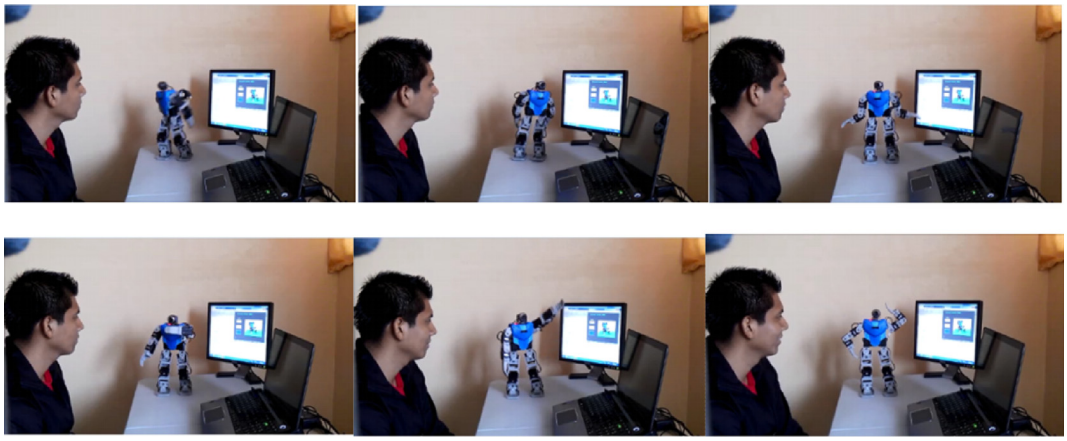


Fig. 17. Dialogue and execution of Routine 4.

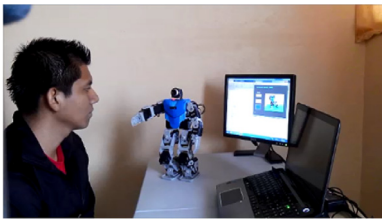


Fig. 18. Dialogue and execution of Routine 5.

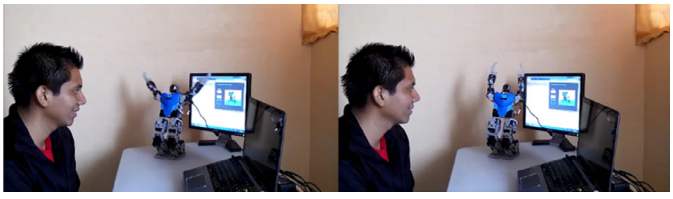


Fig. 19. Dialogue and execution of Routine 2 with different emotion.

MX-Expressions databases. Therefore, the development of the multimodal system which considered the use of evolutionary computation represents an important technological advance to achieve more efficient human-robot interaction systems.

Among the strengths of the present work the following can be mentioned:

- the multimodal system is based of standard techniques (HMMs, PCA, ANN) and the specific structures that make these techniques appropriate for the problem are determined by means of GAs;
- different training-testing schemes (cross-validation) support the consistent performance of the multimodal system;
- performance of the speech and vision systems is consistent independently of the user's gender;
- the rules for phonetic transcription of Mexican speech can provide suitable training data for recognition of emotional speech.

However there is a limitation of the present work which is related to its comparison with other works. As such, there is not direct comparison due to the specific databases created for this work which were focused on Mexican users. Also, the main performance was assessed on live tests. These limitations are considered as main points to be addressed in future work:

- Increase the recognition rate of the speech system under the speaker-independent scheme.
- Consider other HMM structures for optimization with GA.
- Perform testing of the vision system with other databases such as FEEDTUM and FACES. Compare recognition performance with these databases and other recognition techniques.
- Make a comparison between PCA and Fisher for the extraction of features that best discriminate classes instead of extracting features that best describe a class.
- Use other optimization techniques as Tabu Search and Ant Colony.
- Consider other speech coding techniques for feature extraction of emotional speech.
- Make use of depth cameras for facial expression modeling in the three-dimensional space.
- Develop a multimodal system with support of psychology to verify and change the emotions of human users.
- Establish the use of humanoid robotic systems in centers for therapies and care of adult and young people.

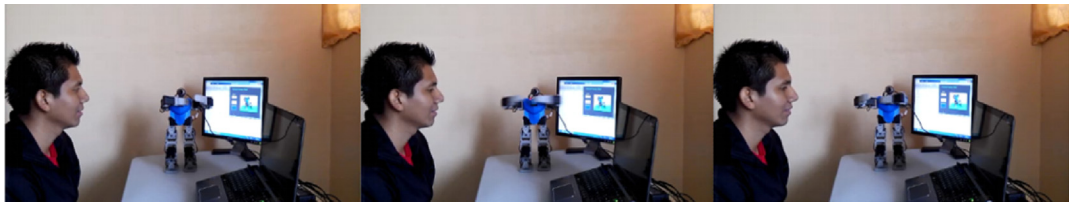


Fig. 20. Dialogue and execution of Routine 5 with different emotion.

References

Ali, H., Hariharan, M., Yaacob, S., & Hamid, A. (2015). Facial emotion recognition using empirical mode decomposition. *Expert Systems with Applications, 42*(3), 1261–1277.

Alter, K., Rank, E., & Kotz, S. A. (2000). Accentuation and emotions - two different systems? In *Proceedings of the 2010 ISCA Workshop on Speech and Emotion, ITRW '00: 1* (pp. 138–142).

Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review, 43*, 155–177.

AT&T Technology Solutions (2015). AT&T FSM Library. In http://www.research.att.com/export/sites/att_labs/library/documents/licensing_data_sheets/fsmlibrary_factsheet_20090925.pdf Accessed in 30/06/2015.

Austermann, A., Esau, N., Kleinjohann, L., & Kleinjohann, B. (2005). Fuzzy emotion recognition in natural speech dialogue. In *Proceedings of the 2005 IEEE International Workshop on Robot and Human Interactive Communication, ROMAN '05* (pp. 317–322).

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Archy, S., Russell, M., & Wong, M. (2004). "you stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC '04* (pp. 171–174).

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies, 59*, 119–155.

Busso, C., Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Deng, Z., Lee, S., Neumann, U., & Narayanan, S. (2004). Analysis of Emotion Recogition Using Facial Expressions, Speech and Multimodal Information. In *Proceedings of the International Conference on Multimodal Interfaces, ICMI '04* (pp. 205–211).

Caballero, S. (2013). Recognition of emotions in Mexican Spanish speech: An approach based on acoustic modelling of emotion-specific vowels. *The Scientific World Journal, 2013*, 1–13.

Chaavan, V. M., & Gohokar, V. V. (2012). Speech emotion recognition by using SVM-classifier. *International Journal of Engineering and Advanced Technology, 1*(5), 11–15.

Chambers S., Breazel C., Atkins A., Revis M., Asher J., Craft A., Westelman R., Kotelly B., & Smith L. (2015). Meet Jibo, The World's First Family Robot. In http://www.jibo.com/ Accessed in 30/06/2015.

Chaturvedi, A., & Tripathi, A. (2014). Emotion recognition using fuzzy rule-base system. *International Journal of Computer Applications, 93*(11), 25–28.

Chen, L., Mao, X., Xue, Y., & Cheng, L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing, 22*, 1154–1160.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing, 28*(4), 357–366.

Filko, D., & Martinovic, G. (2013). Emotion recognition system by a neural network bases facial expression analysis. *Automatika, 54*(2), 263–272.

Firoz-Shah, A., Vimal-Krishnan, V. R., Raji-Sukumar, A., Jayakumar, A., & Babu-Anto, P. (2009). Speaker independent automatic emotion recognition from speech: A comparison of MFCCs and discrete wavelet transforms. In *Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing, ARTCOM '09* (pp. 528–531).

Gosavi, A. P., & Khot, S. R. (2013). Facial expression recognition using principal component analysis. *International Journal of Soft Computing and Engineering, 3*(4), 258–262.

Haq, S., Jackson, P. J. B., & Edge, J. (2008). Audio-visual Feauture Selection and Reduction for Emotion Classification. In *Proceedings of Auditory-Visual Speech Processing, AVSP '08* (pp. 185–190).

Ilbeygi, M., & Hosseini, H. (2012). A novel fuzzy facial expression recognition system based on facial feature extraction from color face images. *Engineering Applications of Artificial Intelligence, 25*, 130–146.

Karthigayan, M., Rizon, M., Nagarajan, R., & Sazali, Y. (2008). Genetic Algorithm and Neural Network for Face Emotion Recognition. In *Affective computing* (pp. 57–68). InTech.

Kaur, M., Vashisht, R., & Neeru, N. (2010). Recognition of facial expressions with principal component analysis and singular value decomposition. *International Journal of Computer Applications, 9*(12), 36–40.

Kulic, D., & Croft, E. A. (2007). Affective state estimation for human-robot interaction. *IEEE Transactions on Robotics, 23*(5), 991–1000.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., & Narayanan, S. (2004). Emotion Recognition based on Phoneme Classes. In *Proceedings of the 2004 International Conference on Spoken Language Processing, ICSLP '04: 1* (pp. 889–892).

Li, A., Fang, Q., Hu, F., Zheng, L., Wang, H., & Dang, J. (2010). Acoustic and articulatory analysis on Mandarin Chinese vowels in emotional speech. In *Proceedings of the 7th International Symposium on Chinese Spoken Language Processing, ISCSLP '10* (pp. 38–43).

Lisetti, C., Nasoz, F., LeRouge, C., Ozyer, O., & Alvarez, K. (2003). Developing multimodal intelligent affective interfaces for tele-home health care. *International Journal of Human-Computer Studies, 59*, 245–255.

López, J., Cearreta, I., Garay, N., López, K., & Beristain, A. (2006). Creación de una base de datos emocional bilingüe y multimodal. In M. A. Redondo, C. Bravo, & M. Ortega (Eds.), *Proceedings of the 7th Spanish Human Computer Interaction Conference, Interaccion '06* (pp. 55–66).

Luo, Y., Wu, C., & Zhang, Y. (2013). Facial expression recognition based on fusion feature of PCA and LBP with SVM. *Optik - International Journal for Light and Electron Optics, 124*(17), 2767–2770.

Mohri, M., Pereira, F., & Riley, L. (1996). Weighted Automata in Text and Speech Processing. In *Proceedings of the 12th European Conference on Artificial Intelligence, ECAI '96* (pp. 257–286).

Odashima, T., Onishi, M., Riken, N., Hirano, S., Mukai, T., & Luo, Z. (2008). Development of the tactile sensor system of a human-interactive robot "RI-MAN". *IEEE Transactions on Robotics, 24*(2), 505–512.

Owusu, E., Zhan, Y., & Mao, Q. R. (2014). A neural-AdaBoost based facial expression recognition system. *Expert Systems with Applications, 41*(7), 3383–3390.

PAL Robotics (2015). Reem - humanoid robot. In http://reemc.pal-robotics.com/en/ Accessed in 30/06/2015.

Pal, S. S., & Hasan, M. (2014). Facial expression recognition using fuzzy logic. *International Journal of Science and Research, 3*(6), 851–854.

Pao, T. L., Liao, W. Y., Chen, Y. T., Yeh, J. H., Cheng, Y. M,, & Chien, C. S. (2007). Comparison of several classifiers for emotion recognition from noisy Mandarin speech. In *Proceedings of the 3rd International Conference on International Information Hiding and Multimedia Signal Processing, IIHMSP '07)* (pp. 23–26).

Pérez-Gaspar, L., Caballero-Morales, S. O., & Trujillo-Romero, F. (2015a). Factores en el Reconocimiento Facial de Emociones y la Integración de Optimización Evolutiva. *Research in Computing Science, 91*, 45–56.

Pérez-Gaspar, L., Caballero-Morales, S. O., & Trujillo-Romero, F. (2015b). Integración de Optimización Evolutiva para el Reconocimiento de Emociones en Voz. *Research in Computing Science, 93*, 9–21.

Pineda, L., Villaseñor, L., Cuétara, J., Castellanos, H., Galescu, L., Juárez, J., Llisterri, J., & Pérez, P. (2010). The corpus DIMEX100: Transcription and evaluation. *Language Resources and Evaluation, 44*, 347–370.

Pooja, R. N., & Kaur, S. (2010). Hybrid technique for human face emotion detection. *International Journal of Advanced Computer Science and Applications, 1*(6), 91–101.

Prolific Technology (2015). USB-TTL serial converter datasheet. http://www.prolific.com.tw/userfiles/files/ds_pl2303hxd_v1_4_4.pdf Accessed in 30/06/2015.

Rao, K. S., Saroj, V. K., Maity, S., & Koolagudi, S. G. (2011). Recognition of emotions from video using neural networks models. *Expert Systems with Applications, 38*(10), 13181–13185.

Rasoulzadeh, M. (2012). Facial expression recognition using fuzzy inference system. *International Journal of Engineering and Innovative Technology, 1*(4), 1–5.

Robotis (2012). *Bioloid Premium, Quick Start: Assembly and Program Download Manual*. Robotis Co., Ltd..

Röfer, T., Laue, T., Burchardt, A., Damrose, E., Fritsche, M., Müller, J., & Rieskamp, A. (2008). B-Human: Team Description for RoboCup 2008. In Iocchi L., Matsubara H., Weitzenfeld A., & C. Zhou (Eds.), *Pre-proceedings of Robocup 2008: Robot Soccer World Cup XII* (pp. 1–6).

Samani, H. A., & Saadatian, E. (2012). A multidisciplinary artificial intelligence model of an affective robot. *International Journal of Advanced Robotic Systems, 9*, 1–11.

Schuller, B., Muller, R., Hornler, B., Konosu, H., & Rigoll, G. (2007). Audiovisual Recognition of Spontaneous Interest within Conversations. In *Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI '07* (pp. 30–37).

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *Proceedings of the 2003 International Conference on Multimedia and Expo, ICME '03: 2* (pp. 401–404).

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing, 1*(2), 119–131.

Shin, Y.-G., Park, S. S., Kim, J.-N., Kim, J.-N., & Jang, D.-S. (2006). Development of a Humanoid Robot for Emotion Recognition. In *Proceedings of the 5th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, CIMMACS '06* (pp. 308–314).

Shing, C., Phooi, K., Ang, L.-M., & Wern, L. (2014). A new approach of audio emotion recognition. *Expert Systems with Applications, 41*(13), 5858–5869.

Song, M., Bu, J., Chen, C., & Li, N. (2004). Audio-Visual-Based Emotion Recognition: A New Approach. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition, CVRP '04: 2* (pp. 1020–1025).

Song, M., You, M., Li, N., & Chen, C. (2008). A robust multimodal approach for emotion recognition. *Neurocomputing, 71*, 1913–1920.

Tayal, S., & Vijay, S. (2012). Human emotion recognition and classification from digital colour images using fuzzy and PCA approach. *Advances in Computer Science, 167*, 1033–1040.

Thuseethan, S., & Kuhanesan, S. (2014). Eigenface based recognition of emotion variant faces. *Computer Engineering and Intelligent Systems, 5*(7), 31–37.

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision, 57*(2), 137–154.

Vlasenko, B., Schuller, B., Wendemut, A., & Rigoll, G. (2007). Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In *Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, ACII '07: 4738* (pp. 139–147). Springer Berlin Heidelberg.

Wan, Z., & Guan, L. (2005). Recognizing Human Emotion from Audiovisual Information. In *Proceedings of the 2005 International Conference on Acoustics, Speech and Signal Processing, ICASSP '05* (pp. 1125–1128).

Wu, C. H., & Liang, W. B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing, 2*(1), 10–21.

Young, S., & Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

Yu, F., Chang, E., Xu, Y. Q., & Shum, H. Y. (2001). Emotion detection from speech to enrich multimedia content. In *Advances in Multimedia Information Processing, Lecture Notes in Computer Science: Proceedings of the 2nd IEEE Pacific-Rim Conference on Multimedia, PCM '01: 2195* (pp. 550–557). Springer Berlin Heidelberg.

Yu, W. (2008). Research and implementation of emotional feature classification and recognition in speech signal. In *Proceedings of the 2008 International Symposium on Intelligent Information Technology Application Workshops, IITAW '08* (pp. 471–474).

Yun, S., & Yoo, C. D. (2009). Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '09* (pp. 4169–4172).

Zavaschi, T., Britto, A., Oliveira, L., & Koerich, A. (2013). Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications, 40*(2), 646–655.

Zhang, L., Jiang, M., Farid, D., & Hossain, M. A. (2013). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications, 40*(13), 5160–5168.

Zhang, Y., Zhang, L., & Hossain, M. A. (2015). Adaptive 3D facial action intensity estimation and emotion recognition. *Expert Systems with Applications, 42*(3), 1446–1464.