In Search of Primary Rubrics for Language Independent Emotional Speech Identification

R. Barra, J. Macías-Guarasa, J.M. Montero, C. Rincón, F. Fernández and R. Córdoba
Speech Technology Group. Department of Electronic Engineering
Universidad Politécnica de Madrid, Spain
{barra, macias, juancho, carmenr, efhes, cordoba}@die.upm.es

***Abstract* – In this paper we describe recent work oriented towards establishing some background to support the design of successful "universal" (language independent, specifically) emotional speech identification systems. Our final goal is using such systems to improve state of the art advanced spoken language systems.**

We carried out several experiments with emotional speech databases for Spanish (SES) and German (EMODB), including language-dependent tasks (to establish the accuracy of the identification systems), and language independent (cross-language) tasks.

We show how automatic emotion identification results can be comparable to those obtained by human listeners, provided that enough training data is available. Moreover, and in spite of data availability, we also show how there are emotions (sadness and anger) that are clearly identified in the automatic experiments, for both languages, showing a reasonably clear "language-independent" behavior of such emotions.

Finally, we show how psychological considerations related to the emotional speech generation process can be somehow related to our experimental results, suggesting that previously identifying such psychological features may greatly help the development of high quality "universal" emotional speech identification systems.

***Keywords* – emotion identification, human-machine interfaces**

I. INTRODUCTION

The idea of making artificial systems sensitive to a user's emotional state offers promise in improving the state of the art in spoken dialog systems [1].

As a part of an effort to make spoken dialog systems increasingly natural, it is beneficial to enable these systems not only to recognize a particular sequence of words or the content encoded in a user's response; but also to extract information about the emotional state of the user. This additional information can then be used by the dialog management framework for a number of highly natural response behaviors, such as to avoid misunderstandings and recovering from errors. Additionally, information about a user's emotional state has potential specific uses in certain task domains such as tutoring [2], tele-marketing and health counseling [3].

In goal-oriented human machine communication the user may display different states due to prior conditions (e.g. previous attempts at solving the task), poor performance, poor machine cooperativeness in acknowledging and/or solving a problem (e.g. machine misunderstanding) or poor reward (e.g. the dialog is not successful) [4]. In this context, the identification of the emotional state of the user is essential for advanced behavior of the spoken dialog system.

In addition to all the above, and given the increasingly global world we live in, multilingual human-machine dialog systems are being increasingly demanded. In this context, a language-independent emotional speech classification system would be extremely useful. In the literature there are very few examples of previous work related to language-independent emotional speech identification (such as in [5], using a neural network approach), although there is very little insight in providing actual theoretical and empirical clues to give significant support on the apparent adequacy of their experimental results to the language-independent nature of the required systems. In our work, we have reviewed the literature regarding psychological basis of human emotional speech and will confront the existing frameworks with the actual results obtained in our experiments.

To briefly summarize, the emotional state is often modelled in psychology science using three abstract dimensions, arousal or appraisal (related with a physiological and psychological state that involves the activation of several biological systems), valence (related with the positive or negative reactions to different events, reactions and objects) and dominance (related with the organism's estimation of how well it would be able to to cope with a particular stimulus event and its consequences)

This work has been partially supported by EDECAN (TIN2005-08660-C04-04), ROBINT (DPI2004-07908-C02-02), ATINA (UPM-CAM REF: CCG06-UPM/COM-516) and TINA (UPM-CAM REF: R05/10922).

[6]. As pointed in [7], basic emotions, named also as "primary emotions" in the literature, are said to be common to all humans, outperforming races, cultures and languages. Mapping primary emotions into this dimensional space and its identification by automatic detection of each dimension values [4] would perform the emotional state identification of a generic (language-independent) speaker.

The remainder of this work is organized as follows. First, the corpora evaluated is described. Next, the experimentation methodology is explained and results from two different language-dependent emotion classifiers (Spanish and a German emotions classifier) are presented. Continuing, results from a language-independent emotions identification experiment are shown. Getting inside emotions nature, some primary emotion rubrics common to both languages are analyzed; and the relation with their projections onto the arousal-valence-dominance space. Finally, the paper concludes summarizing the principal results of the investigation.

II. Corpora description

In emotion research, it is typical to use acted emotional recordings instead of natural data, even that there are many arguments in disfavor of acted emotional expression. The main problem is that full-blown emotions rarely appear in the real world [8] and, furthermore, there are physical emotional cues that cannot be consciously mimicked.

However, as clear emotional expressions are not only rare in everyday situations but also the recording of people experiencing full-blown emotions is ethically problematic, it is almost impossible to use natural data if basic emotions are the subject of investigation.

When designing the experimental setup, we also wanted to stress the language-independent characteristic of our search for primary rubrics, so that the selection of languages was aimed at choosing two (available) languages with a different historical origin and evolution.

This is the reason why, in this work, we have used two different language databases, for Spanish and German, both of them comprising acted emotional speech.

A. Spanish corpus

The Spanish Emotional Speech corpus (SES), described in [9], contains three emotional speech recording sessions played by a professional male actor in an acoustically-treated studio. Each recorded session includes thirty words (2 minutes), fifteen short sentences (7 minutes) and four paragraphs (39 minutes), simulating three basic or primary emotions (sadness, happiness and cold anger), one secondary emotion (surprise) and a neutral speaking style. The text uttered by the actor did not convey any explicit emotional content.

This parallel corpus was phonetically labeled in a semiautomatic way. An automatic pitch epoch extraction software was used, but the outcome was manually revised using a graphical audio-editor program, which was also used for phoneme location and labeling.

The assessment of the emotional voice was aimed at evaluating the speech corpus as a model for recognizable emotional speech [9]. Perceptual copy-synthesis tests (mixing emotional phoneme durations and linearized F0 contours with neutral diphones or vice versa), showed the basic segmental or non segmental nature of each emotion, as described in [10].

Emotional patterns were also evaluated by means of automatic identification experiments in [11]. Emotional information was analyzed using segmental (MFCC) and prosodic information (F0-related statistics). When both sources of information were combined, better classification rates were achieved, even for mainly "prosodic" emotions.

B. German corpus

The Berlin Database of Emotional Speech (EMODB) is an emotional German speech corpus fully described in [12]. Ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and are interpretable in all applied emotions.

The utterances were recorded in an anechoic chamber with a high-quality recording equipment. In addition to the speech signal, electro-glottograms were also recorded. The speech material comprises around 24.5 speech minutes recorded in 800 sentences (seven emotions * ten actors * ten sentences + some second versions), but the amount of speech data per emotion is very small, almost 4 minutes per emotion.

The complete database was evaluated in a perceptual test regarding the recognizability of emotions and their naturalness, as shown in [12]. Utterances recognized better than 80% and judged as natural by more than 60% of the listeners were phonetically labelled in a narrow transcription with special markers for voice-quality, phonatory and articulatory settings and articulatory features.

III. Experiments

Given our goal of finding multi-lingual primary rubrics for emotion identification, in this work, a set of identification experiments have been run.

First of all, we have separately analyzed the emotions of each corpus, by means of an automatic emotion identification system. Second, a comparison between the automatic (objective) identification rates and their corresponding perceptual evaluation has been done. Finally, the similarities of the emotions between both languages has been studied, in order to find relevant features to be further exploited in multi-lingual human-machine spoken dialog systems.

All utterances were processed by a frame analysis using a 25ms window with 10ms frame shift. Mel Frequency cepstrum Coefficients (MFCC) were extracted to represent the speech signal from a segmental point of view. MFCCs have

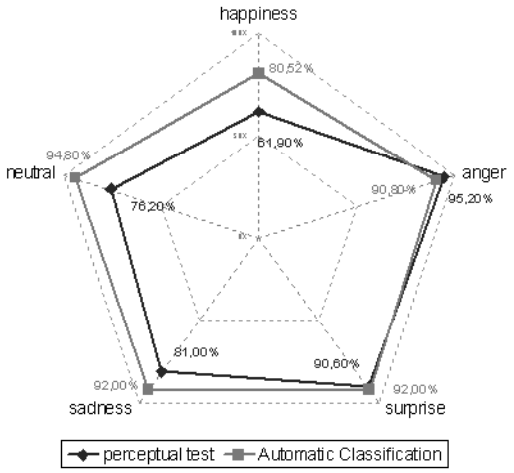

Fig. 1. IDENTIFICATION RATES FOR SES EMOTIONS BY HUMAN LISTENERS AND AUTOMATIC CLASSIFIER.

TABLE II.
CONFUSION MATRIX FOR EMODB EMOTIONS (FIGURES IN %).

	no Norm.	CMN	CVN	CMN+CVN
Identification Rate	42.18	48.08	46.09	51.33
Relative gain		14	9.3	21.7

TABLE I.
CONFUSION MATRIX FOR SES EMOTIONS (FIGURES IN %).

ACTED EMOTION	IDENTIFIED EMOTION				
	happiness	anger	surprise	sadness	neutral
happiness	80.5	2.3	14.9	1.1	1.1
anger	2.3	90.8	2.3		4.6
surprise	6.9	1.1	92		
sadness		1.1		92	6.9
neutral	3.4			1.7	94.8
PRECISION	86.5	95.3	84.3	97	88.3

demonstrated to achieve reasonable performance in similar emotion identification tasks as described in [11].

For identification purposes, every emotion was modeled using a Gaussian Mixture Model (GMM), with a low number of distributions (given that initial experiments with both corpus show lack of data to train models with a high number of gaussians).

A Bayes classifier was used to implement our emotion recognizer, in which the probability of a certain emotion given the acoustic evidence is calculated by accumulating the probabilities of the frames corresponding to each utterance.

A. Spanish emotions identication

The behavior of a real-world spoken dialog system could be improved if the user's emotional state could be estimated along several dialog turns. Since the utterances in SES have no emotional context, an approach to this scenario was built by trying to identify the intended emotion intended trough several sentences, which corresponds to the SES paragraphs. In this experiment, the emotional models were trained with SES sentences and the SES paragraphs were used for testing. The average error emotion identification rate in this task was 97.62%, showing the adequacy of the identification system for the task.

An additional experiment was carried out by training the system with the SES paragraphs and testing with the SES sentences, leading to a scenario in which emotion identification would be decided for every single sentence (in a given interaction with the spoken dialog system). In this (more difficult task), the confusion matrix of Table I summarizes the achieved results.

Finally, a comparison between automatic clasification and perceptual subjective identication rates by human listeners was carried out ([9]). Its results are displayed in Figure A, showing that the automatic system lead to similar identification rates (89.98% in average) than the ones obtained by human listeners (90.2% in average).

B. German emotions identication

In the case of the EMODB, only short utterances where available for the classification task. Due to the existence of several speakers, and the lower ammount of data, a cross-validation strategy has been adopted. Nine of the ten speakers were used to train the system and the other one was used for testing. The experiment was run classifying one of the speakers each time. The average final identification rate was 42.18%, much lower than in SES.

Our strategy to improve this result uses well known normalization techniques (widely used in speech recognition or speaker identification tasks). We carried out experiments using Cepstral Mean Normalization (CMN), Cepstral Variance Normalization (CVN) and its combination (CMN+CVN), significantly reducing the emotion identification error rates. Table II shows the results applying each normalization scheme and the corresponding relative gain in performance.

As can be clearly seen, CVN+CMN is the normalization technique that presents the highest performance (51.33%, with a 21.7% relative performance gain), but still at relatively low identification rate (although well above chance). The confusion matrix obtained using this technique is shown in Table III.

As in the identification experiments of SES emotions, the comparison between the results obtained in the perceptual evaluation of the corpus [12] and the automatic clasification are plotted in Figure B. The average identification rate achieved by human listeners (86.08%) clearly outperforms the results obtained by the automatic classifier (51.33%), which is also much lower than the figures for the Spanish SES database. A careful study on this poor performance, reveals that the ammount of data is a major problem in EMODB. In spite of this important issue, some emotions in EMODB were identified with high accuracy, with results similar to the ones obtained by human listeners, such as disgust (humans 79.60% vs. automatic 73.91%), sadness (80.70% vs. 95.16%) and anger (96.90% vs. 71.65%).

TABLE III.
CONFUSION MATRIX FOR EMODB EMOTIONS (FIGURES IN %).

ACTED EMOTION	IDENTIFIED EMOTION						
	happiness	anger	boredom	sadness	fear	disgust	neutral
happiness	42.25	26.76			23.94	5.63	1.41
anger	13.39	71.65			14.17	0.79	
boredom			43.21	12.35	6.17	12.35	25.93
sadness			1.61	95.16			3.23
fear	8.70	13.04	4.35	7.25	44.93	11.53	10.14
disgust	2.17		2.17		13.04	73.91	8.79
neutral			36.71	3.80	5.06	11.39	43.04
PRECISION	63.52	64.29	49.07	80.26	41.87	63.90	46.55

TABLE IV.
CONFUSION MATRIX FOR CROSS-LANGUAGE EMOTIONS (FIGURES IN %).

ACTED EMOTION (EMODB)	IDENTIFIED EMOTION (SES models)			
	happiness	anger (cold)	sadness	neutral
happiness	77.46	21.13		1.41
anger (hot)	77.17	22.83		
sadness			100	
neutral	12.66	10.13	41.77	35.44
PRECISION	46.31	42.22	70.54	96.18

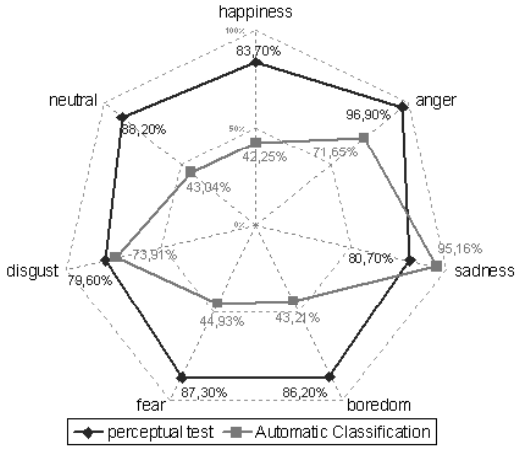

Fig. 2. IDENTIFICATION RATES FOR EMODB EMOTIONS BY HUMAN LISTENERS AND AUTOMATIC CLASIFIER.

C. Cross-language emotions identification

The nature of basic emotions [13], named as *primary emotions*, is supposed to present common rubrics in human beings, independently of their culture or language. Properly modelling these emotions could contribute to broad personalization of user profiles in spoken dialog systems.

We approach this "language-independent" scenario by training the automatic classifier with emotional speech from one of the languages and classifying the emotional speech in the other language. Given the databases used, only the common emotions between both corpus can be tested. In our case, the Spanish database was selected to train the system, mainly due to the availability of a higher amount of data and because similar results were obtained in SES when comparing automatic classification with perceptual tests (as an indication of the higher similarity between the human and machine ability to detect emotions for Spanish). In this case, the acoustic features were normalized using the CMN+CVN technique.

Table IV presents the confusion matrix in this cross-language experiment (train with Spanish emotions and test with German emotions). The results show two of the emotions as highly identifiable, in a "language-independent" way: happiness (77.46%) and sadness (100%), leading to some hope for this claim of the existence of common universal rubrics in emotional speech.

IV. DISCUSSION

A. Spanish emotions identification

The emotional patterns used by the Spanish actor were clearly identified by Spanish listeners and by our automatic classifier (humans 90.2% vs. automatic 89.98%, in average). For both tests (humans and automatic), the less identifiable emotion was happiness (humans 61.9% vs. automatic 80.5%). The Spanish emotions with lowest arousal level, sadness and neutral, are confused. Sadness is confused 6.9% with neutral. Similar results are obtained in emotions with positive valence, like happiness and surprise, that tend to be confused by the classifier. Happiness is confused with surprise (14.9%) and surprise is also confused with happiness (6.9%). Additionally, these two positive emotions have the lowest precisions (86.5% and 84.3% respectively).

B. German emotions identification

For the German EMODB database, normalization techniques, often used in automatic speech recognition, showed consistent improvements; leading to 21.7% of relative performance gain when CMN+CVN is applied to subtract channel and inter-speaker variability to the emotions acted by German speakers. The lower identification rates suggests that more data would be necessary to achieve results similar to the ones obtained by German listeners (86.58%).

German sadness is clearly identified by the automatic classifier (95.16%), even better than German listeners (80.70%).

TABLE V.
CONFUSION MATRIX FOR CROSS-LANGUAGE EMOTIONS (FIGURES IN %).

ACTED AROUSAL LEVEL	IDENTIFIED AROUSAL LEVEL	
	high	low
high	99.49	0.51
low	12.77	87.23
PRECISION	88.63	99.22

Good results are also achieved for anger (71.65%) and disgust (73.9%).

As in Spanish emotions identification, low arousal emotions (like sadness, boredom or even neutral and disgust) are confused one each others. Boredom is identified as sadness (12.35%), disgust (12.35%) and neutral (25.93%). Sadness is only confused with boredom (1.61%) and neutral (3.23%). Neutral is high identified with disgust (11.39%) and boredom (36.71%).

Same confusion is obtained between (high arousal emotions) happiness, hot anger and fear. Happiness is specially confused with anger (26.76%) and fear (23.94%). Anger is also confused with happiness (13.39%) and fear (14.17%). Fear, with low identification rate (44.93%), is specially confused with anger (13.04%). Sadness, both in Spanish and German is the emotion with the highest precision in both languages (97% and 80.26% respectively).

C. Cross language emotions identification

Cross-language results (Table IV) present a common set of primary rubrics for common emotional behaviors in both corpora.

German sadness is completely identified when using the Spanish sadness model. Also German happiness is highly identified (77.46%). On the contrary, German anger is confused with happiness (77.17%) instead of with Spanish anger (22.83%). This can be explained due to the different nature of both angers. German anger (described as hot anger) has low dominance and high arousal levels, while Spanish anger (described as cold anger) has a high dominance level and a reduced arousal level.

To get further insight in this behavior, we proyected the common set of emotions into a single dimension reflecting the arousal level (high arousal for happiness and anger, versus low arousal for neutral and sadness). In this case, a very high identification rate of the arousal level could be obtained 99.49% for high arousal and 87.23% for low arousal, as shown in Table V.

V. CONCLUSION

In this work, several identification experiments of acted emotions in two different languages were done, with the objective of establishing some background to support the design of successful "universal" (language independent, specifically) emotional speech identification systems. Our final goal is using such systems to improve state of the art advanced spoken language systems.

First of all, we showed the importance of applying speech normalization strategies when dealing with cross-speaker and, more important, cross-language emotional speech identification tasks, achieving improvements of up to 21.7% for CMN+CVN.

We also showed how automatic emotion identification results are comparable to those obtained by human listeners, provided that enough training data is available (such as in the Spanish SES database). Moreover, and in spite of data availability, there are emotions (sadness and anger) that are clearly identified in the automatic experiments, for both languages, showing a reasonably clear "language-independent" behavior of such emotions.

High identification rates for sadness (100%) and happiness (77.46%) has been shown in a cross-language experiment (training with Spanish and testing with German). Sadness presents a clear emotional pattern, due to its persistent high precision in all the experiments.

Regarding psychological considerations related to emotional speech generation, valence, arousal and dominance features have been also been observed by the identification experiments. Spanish emotions with positive valence level (happiness and surprise) are somehow confused. Similar considerations can be extracted for the dominance level of cold anger (high dominance) and hot anger (low dominance). High arousal emotions (happiness, anger or fear) are confused between them; same as emotions with low arousal level (boredom, sadness, neutral or disgust). Additionally, high performance detection of arousal level of language-independent emotions is implemented.

All these results suggest that in the implementation of a high quality emotional speech identification system, previously identifying the dominance, valence and, specially, arousal level of spoken utterances in a real spoken dialog system, would help to identify the emotional state of the user regardless to the user's language. This idea is also supported by studies such as [4], in which a simple binary representation of emotions (positive vs. negative) suffices in the context of a goal-driven conversational system.

REFERENCES

[1] Rohit Kumar, Carolyn P. Rose, and Diane J. Litman, "Identification of confusion and surprise in spoken dialog using prosodic features," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh, PA., USA, 2006, pp. p. 1921–Wed2BuP.14.

[2] K. Forbes-Riley D. Litman, "Recognizing emotions from student speech in tutoring dialogs," in *Proceedings of IEEE ASRU Workshop 2003*, Pittsburgh, PA., USA, 2003.

[3] T. Bickmore, *Relational Agents: Effecting changes through Human Computer Relationships*, PhD thesis, MIT Media Arts and Science, 2003.

[4] G. Riccardi and D. Hakkani-Tur, "Grounding emotions in human-machine conversational systems," *Lecture notes in computer science*, , no. 3840, pp. 114–154, 2005.

[5] Muhammad Waqas Bhatti, Yongjin Wang, and Ling Guan, "A neural network approach for human emotion recognition in speech," in *Proceedings of the 2004 IEEE International Symposium on Circuits and Systems*, Vancouver, British Columbia, Canada, 2004, vol. 5, pp. 181–184.

[6] Marc Schoder, *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*, PhD thesis, Institut fur Phonetik. Universitat des Saarlandes, Saarbruken, Germany, Juni 2004.

[7] R.R. Cornelius, "Theoretical approaches to emotion," in *Proceedings of the ISCA Workshop on Speech and Emotion*, Northern Ireland, Sep. 2000, pp. 3–30.

[8] E. Douglas-Cowie, R. Cowie, and M.I. Schroder, "A new emotion database: Considerations, sources and scope," in *Proceedings of the ISCA Workshop on Speech and Emotion*, NewCastle, United Kindom, Sep. 2000.

[9] J.M. Montero, J. Gutirrez-Arriola, S. Palazuelos, E. Enrquez, and J.M. Pardo, "Spanish emotional speech from database to tts," in *Proceedings of ICSLP*, Sep. 1998, pp. 923–925.

[10] J.M. Montero, J. Gutierrez-Arriola, R. Cordoba, E. Enriquez, and J.M. Pardo, "The role of pitch and tempo in emotional speech," in *Improvements in speech synthesis. Ed. Wiley and Sons*, Sep. 2002, pp. 246–251.

[11] R. Barra, J.M. Montero, J. Macias-Guarasa, L.F. DHaro, R. San-Segundo, and R. Cordoba, "Prosodic and segmental rubrics in emotion identi?cation," in *Proceedings of ICASSP*, Sep. 2006, pp. 1085–1088.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of Interspeech*, Lissabon, Portugal, Sep. 2005.

[13] Roddy Cowie and Randolph R. Cornelius, "Describing the emotional states that are expressed in speech," *Speeh Communication*, vol. 1, no. 40, pp. 5–32, 2003.