2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

The Mexican Emotional Speech Database (MESD): elaboration and assessment based on machine learning*

Mathilde M. Duville, Luz M. Alonso-Valerdi, and David I. Ibarra-Zarate

***Abstract*—The Mexican Emotional Speech Database is presented along with the evaluation of its reliability based on machine learning analysis. The database contains 864 voice recordings with six different prosodies: anger, disgust, fear, happiness, neutral, and sadness. Furthermore, three voice categories are included: female adult, male adult, and child. The following emotion recognition was reached for each category: 89.4%, 93.9% and 83.3% accuracy on female, male and child voices, respectively.**

***Clinical Relevance* — Mexican Emotional Speech Database is a contribution to healthcare emotional speech data and can be used to help objective diagnosis and disease characterization.**

I. Introduction

Acoustic cues of emotional speech production are major predictors of health conditions such as depression [1], autism [2], or schizophrenia [3]. Developments in wireless communication and machine learning engineering led to smart healthcare systems designed to detect pathologies from voice signal analysis without medical visitation. Physiological signals are uploaded to a cloud computer where they can be accessed for subjective (undertaken by a physician) or objective (performed by a computational algorithm) analysis [4]. Objective pathological assessments rely on healthcare big data used for classification of diseases [5]. On the other hand, databases of speech signals must also be used to explore the linguistic and emotional perception that characterizes particular pathological conditions. For instance, the development of validated stimuli for affective prosody may be useful to study the behavioural and neuronal impairments that define the atypical emotional perception of autistics [6], [7]. As emotional expression is shaped by cultural variations [8], databases optimized for the population under study are an urgent need. The aim of this work is to provide a Mexican Emotional Speech Database (MESD) that contains single-word utterances for child, female, and male voices, expressed with six basic emotions: anger, disgust, fear, happiness, neutral, and sadness. Two corpora were created: (corpus A) involved the repetition of 24 words across prosodies and voice categories, and (corpus B) offers utterances of words controlled for linguistic (concreteness, familiarity, and frequency of use), and emotional semantic (valence, arousal, and discrete emotions) dimensions. Researchers, engineers, and physicians can rely on utterances from the corpus that is best appropriate to their needs and experimental conditions.

II. Voice Recordings

A. MESD Word Corpus

Nouns and adjectives were selected from two sources: the single-word corpus of the INTERFACE for Castilian Spanish database [9], hereinafter named *corpus A*; and the Madrid Affective Database for Spanish (MADS), creating *corpus B* [10]. Words from corpus A recurred across emotions and voices (child, male, female). Words from corpus B were selected according to the following criteria: (1) subjective age of acquisition under 9-year-old, (2) emotional semantic rating strictly superior to 2.5 (on a 5-point scale) for 5 particular emotions (anger, disgust, fear, happiness, and sadness), (3) valence and arousal ranging from 1 to 4, or from 6 to 9 for emotional words and greater than 4 but lower than 6 for neutral ones (9-point scale). Finally, (4) emotions were matched as regards 3 linguistic features: concreteness, familiarity, and frequency of use ratings. Scores from males, females, and averaged for all subjects were considered separately.

In sum, MESD corpus included 48 words per emotion (24 from corpus A and 24 from corpus B). That is, 288 single words were used for further utterance by male, female, or child voices.

To control frequency, familiarity, and concreteness ratings, R software[1] was used to run a one-way ANOVA on each parameter separately with emotions as factor. Independence of residuals was assessed by Durbin-Watson test. Normality and homogeneity were assessed by Shapiro-Wilk and Bartlett tests, respectively. In case of non-parametricity, Kruskal-Wallis test was applied. Post-hoc tests were used to statistically assess specific differences (Tukey after ANOVA, Wilcoxon tests with p-value adjustment by Holm method after Kruskal-Wallis). In case of significance, outlier values (i.e., ratings for frequency, familiarity or concreteness outside the range defined by percentiles 2.5 and 97.5) were removed until non-significance was reached. Level of significance was set at $p<0.05$.

*Research supported by the Mexican National Council of Science and Technology (grant reference number: 1061809).

Mathilde M. Duville is working with Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey, N.L., México, 64849 (e-mail: A00829725@itesm.mx).

Luz M. Alonso-Valerdi is working with Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey, N.L., México, 64849 (e-mail: lm.aloval@tec.mx).

David Ibarra-Zarate is working with Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey, N.L., México, 64849 (david.ibarra@tec.mx).

[1] R Foundation for Statistical Computing, Vienna, Austria

B. Participants and Ethical Considerations

Participants were volunteers and non-professional actors: 4 adult males, (mean age = 22.75, SD =2.06), 4 adult females (mean age = 22.25, SD =2.50), and 8 children (5 girls and 3 boys, mean age = 9.87, SD = 1.12). They were included in the study if they had grown up in Mexico in a cultural Mexican environment (Mexican academic education and family environments). Participants were excluded if they presented any pathology that impairs emotional behavior, hearing, or speech, or sickness traits affecting voice timbre. No participant had lived in a foreign country (other than Mexico) more than 2 weeks in the last 4 years. A written informed consent was obtained from all participants and children's parents. Recordings were conducted in accordance with the Declaration of Helsinki and approved on July 14th, 2020 by the Ethical Committee of the School of Medicine of Tecnologico de Monterrey (register number within the National Committee of Bioethics CONBIOETICA 19 CEI 011-2016-10-17) under the following number: P000409-autismoEEG2020-CEIC-CR002.

C. Material and Procedures for Voice Recording

Recordings were carried out in a professional recording studio. A microphone Sennheiser e835 with a flat frequency response (100 Hz to 10 kHz), and a Focusrite Scarlett 2i4 audio interface connected to the microphone with an XLR cable and to a computer were used. Audio files were recorded in the digital audio workstation REAPER (Rapid Environment for Audio Production, Engineering, and Recording), and stored as a sequence of 24-bit with a sample rate of 48000Hz.

Adult and child sessions lasted 1 hour, and 30 minutes, respectively. Each adult uttered words from corpora A and B (48 words per emotion). Four children uttered words from corpus A and 4 children uttered the ones from corpus B (24 words per emotion). The order of corpora was counterbalanced across adult sessions. Emotions were randomly distributed in both adult, and child sessions. After familiarizing with the word-dataset, participants were required to utter each word with the corresponding intended emotional intonation: anger, disgust, fear, happiness, neutral, or sadness. Participants were asked to wait at least 5 seconds between 2 utterances in order to focus before each utterance.

III. Emotion Recognition

Before extracting acoustic features, each word was excerpted from the continuous recording of each session to generate an audio file for each individual word.

A. Acoustic Features Extraction and Data Normalization

Praat and Matlab R2019b were used to extract the features detailed in Table I. The Gaussian distribution of the resulting 30 acoustic features was assessed by Shapiro-Wilk test. Considering the lack of normality, a min-max normalization was applied as described in (1):

$$x_{normalized} = \frac{x - min_k}{max_k - min_k} \quad (1)$$

Where x is the feature to be normalized, max_k is the highest value of acoustic feature vector k and min_k is the lowest.

B. Support Vector Machine (SVM) Predictive Model

Matlab R2019b was used to carry out a supervised learning analysis using a cubic SVM classifier. Hyper-parameters were adjusted to a box constraint level (soft-margin penalty) at 10. The multiclass method (one-vs-one or one-vs-all) and the kernel scale parameters was set to "auto", that is, the algorithm was automatically optimized for both parameters according to the dataset. 77% of the data was used for training, and 23% for validation. A stratified train/test split cross-validation method was used and repeated 10 times. Therefore, data were randomly split before each repetition so that each division (training and validation) presented an equal number of words per emotion. Particularly, training data included 222 observations (37 per emotion), and 66 observations (11 per emotion) were used for validation. Accuracy, recall, precision and F-score were computed in accordance with the resulting confusion matrix [12].

C. Adult Voices

Male and female voices were considered as two separate datasets. The input data for training were the 30 normalized acoustic features extracted from each utterance after a dimensionality reduction based on Principal Component Analysis, explaining the 95% of the variance. A classification

TABLE I. Extracted Acoustic Features for Emotion Recognition

Type	Feature	Description
Prosodic	Fundamental frequency or pitch (Hertz)	Mean and standard deviation over the entire waveform.
	Speech rate	Number of syllables per second.
	Root mean square energy (Volts)	Square root of mean energy.
	Intensity (dB)	Mean and standard deviation over the entire waveform.
Voice quality	Jitter (%)	*Jitter local:* average absolute difference between two consecutive periods, divided by the average period.
		Jitter ppq5: 5-point period perturbation quotient. It is the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period.
	Shimmer (%)	*Shimmer local:* the average absolute difference between the amplitude of two consecutive periods, divided by the average amplitude.
		Shimmer rapq5 : 5-point amplitude perturbation quotient. It is the average absolute difference between the amplitude of a period and the average of the amplitude of it and its four closest neighbors, divided by the average amplitude.
	Mean harmonics-to-noise ratio (dB)	Mean value over the entire waveform of ten times logarithm with base 10 of the ratio between the percentage of the signal composed of harmonics and the percentage of the signal composed of noise.
Spectral	Formants (Hertz)	F1, F2, F3: Mean and bandwidth in center.
	Mel Frequency Cepstral Coefficients	1-13 coefficients.

analysis was conducted on utterances from each actor independently. The final version of the MESD was created by selecting for each emotion the utterances from the actor leading to the highest F-score during validation. This process is described in Fig. 1. A classification analysis was applied on the final 288-utterance dataset.

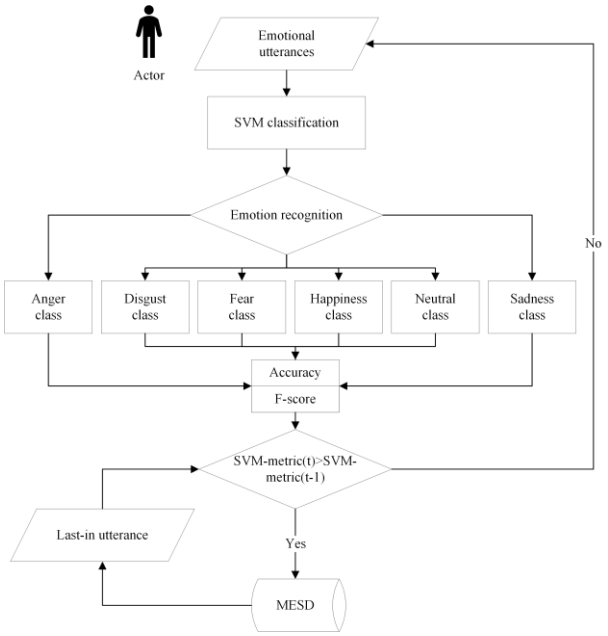

Figure 1. Process to select utterances for adult voices.

D. Child Voices

A k-means clustering analysis was applied on features extracted for each emotion separately (24 observations per participant, leading to 6 datasets of 192 observations). This approach allowed the identification of the highest representative combinations of utterances from actors who uttered corpus A with the ones who uttered corpus B. Namely, it helped to select the most relevant sets of 48 utterances per emotion that will be used as input for the further SVM-based classification. Squared Euclidean distance metric and k-means ++ algorithm for cluster center initialization were used. The optimized number of clusters was assessed by computing silhouette scores. The number of clusters that led to the highest average silhouette score was selected, namely, 2 clusters.

In each cluster, utterances of words coming from corpus A (4 participants) and corpus B (4 participants) were considered separately. For utterances from both corpora, the number of observations for individual participants in each cluster was computed. Pairs of participants (one who uttered words from corpus A and one who uttered word from corpus B) were assessed in each cluster by considering the participant with the highest number of observations. As a result, each pair of participants was composed of 288 utterances (48 per emotion, including 24 of words from each corpus).

Then, a classification analysis was carried out on data from each resulting pair. The input data for training was the 30 normalized acoustic features extracted from each utterance, after a dimensionality reduction based on Principal Component Analysis that explained 95% of the variance. The final version of the MESD was created by selecting for each emotion the utterances from the pair leading to the highest F-score during validation. The resulting set of 288 utterances was used to evaluate the accuracy and F-score for emotion recognition on the final version of MESD.

IV. Results

The MESD is freely available at: http://dx.doi.org/10.17632/cy34mh68j9.1

A. MESD Word Corpus: Corpus B

No inter-emotion difference was emphasized for frequency of use, familiarity, and concreteness ratings after outliers were removed. Namely, for words used for child, male, and female utterances, statistical analysis stressed non-significant p-values ($p>0.05$) for each parameter.

B. Female Adult Voice

Fig. 2 presents the accuracies and F-scores reached for each emotion and their mean values. It is important to note that the most representative female participants for each emotion resulting from the single-actor classification were: (1) participant 2 for anger, (2) participant 2 for disgust, (3) participant 1 for fear, (4) participant 6 for happiness, (5) participant 2 for neutral, and (6) participant 1 for sadness.

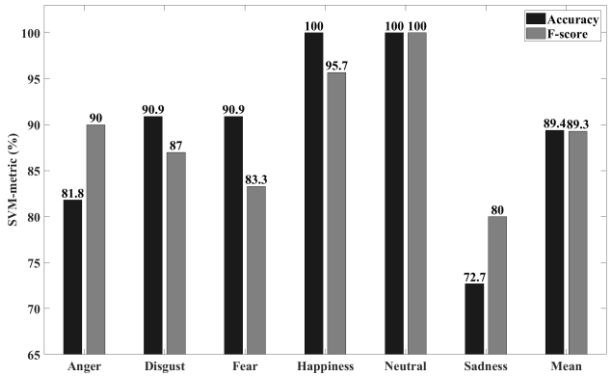

Figure 2. SVM classifier outcome: accuracy and F-score on female voices.

C. Male Adult Voice

Fig. 3 presents the accuracies and F-scores reached for each emotion and their mean values.

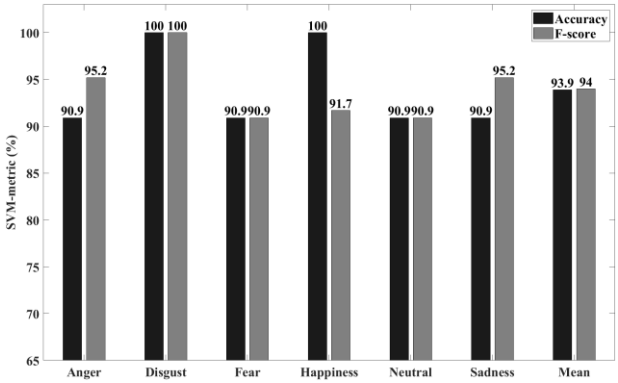

Figure 3. SVM classifier outcome: accuracy and F-score on male voices.

The most representative male participants for each emotion resulting from the single-actor classification were: (1) participant 3 for anger, (2) participant 12 for disgust, (3) participant 3 for fear, (4) participant 3 for happiness, (5) participant 3 for neutral, and (6) participant 12 for sadness.

D. Child Voice

The most representative pairs of child participants for each emotion resulting from the single-pair classification were: (1) participants 16 and 5 for anger, (2) participants 9 and 15 for disgust, (3) participant 16 and 5 for fear, (4) participant 17 and 15 for happiness, (5) participant 16 and 7 for neutral, and (6) participant 16 and 5 for sadness. Fig. 4 presents the accuracies and F-scores reached for each emotion and their mean values.

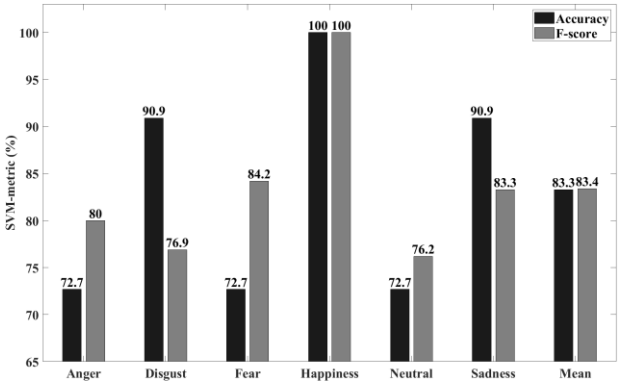

Figure 4. SVM classifier outcome: accuracy and F-score on child voices.

V. Discussion

The MESD contributes to the production of reliable emotional speech data available for healthcare analytics. To date, supplies for linguistic affective stimuli adapted to Mexican Spanish are very scarce [13]. Besides, very few current databases target child voices [14]. The current database presents several advantages: (1) a word corpus controlled for emotional semantic and linguistic parameters was provided [10]; and (2) the MESD includes single-word utterances that contrary to sentences, do not embed variations of emotional information through the utterance [15]. Furthermore, the cognitive processing of emotional words does not involve prediction, integration, and syntactic unification processes that may interplay with the understanding of emotional information [16]. Concreteness, familiarity, and frequency of words from corpus B were controlled to fade trade-off effects between linguistic and emotional processing when using the MESD as stimuli for emotional perception. Nevertheless, words recurrence that characterizes utterances of nouns and adjectives from corpus A may be appropriate for prosodies and voice categories comparisons based on data analysis sensible to phonetic contents. Finally, for both words from corpus A and B, the MESD provides 24 utterances from a unique speaker for each emotional prosody, which guarantees the homogeneity of speaker perception within emotional intonational patterns. As a conclusion, the MESD is a reputable source of emotional utterances that can be applied to (1) big data for smart healthcare, (2) the characterization of normal and pathological emotional prosodies processing and expression, and (3) the exploration of normal or pathological acoustic linguistic information processing and expression.

Acknowledgment

We thank the "Instituto Estatal de la Juventud (INJUVE)", and Norberto E. Naal-Ruiz (https://orcid.org/0000-0002-1203-8925).We acknowledge the Evaluation and Language resources Distribution Agency (ELDA) S.A.S., for sharing the "Emotional speech synthesis database, ELRA catalogue (http://catalog.elra.info), ISLRN: 477-238-467-792-9, ELRA ID: ELRA-S0329".

References

[1] S. Shinohara *et al.*, "Evaluation of the Severity of Major Depression Using a Voice Index for Emotional Arousal," *Sensors*, vol. 20, no. 18, p. 5041, Sep. 2020, doi: 10.3390/s20185041.

[2] D. J. Hubbard, D. J. Faso, P. F. Assmann, and N. J. Sasson, "Production and perception of emotional prosody by adults with autism spectrum disorder: Affective prosody in ASD," *Autism Res.*, vol. 10, no. 12, pp. 1991–2001, Dec. 2017, doi: 10.1002/aur.1847.

[3] D. Chakraborty *et al.*, "Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Apr. 2018, pp. 6024–6028. doi: 10.1109/ICASSP.2018.8462102.

[4] M. Alhussein and G. Muhammad, "Automatic Voice Pathology Monitoring Using Parallel Deep Models for Smart Healthcare," *IEEE Access*, vol. 7, pp. 46474–46479, 2019, doi: 10.1109/ACCESS.2019.2905597.

[5] M. S. Hossain and G. Muhammad, "Healthcare Big Data Voice Pathology Assessment Framework," *IEEE Access*, vol. 4, p. 10, 2016.

[6] J. S. Mulcahy, M. Davies, L. Quadt, H. D. Critchley, and S. N. Garfinkel, "Interoceptive awareness mitigates deficits in emotional prosody recognition in Autism," *Biol. Psychol.*, vol. 146, p. 107711, Sep. 2019, doi: 10.1016/j.biopsycho.2019.05.011.

[7] R. Lindström *et al.*, "Atypical perceptual and neural processing of emotional prosodic changes in children with autism spectrum disorders," *Clin. Neurophysiol.*, vol. 129, no. 11, pp. 2411–2420, Nov. 2018, doi: 10.1016/j.clinph.2018.08.018.

[8] P. Laukka *et al.*, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features.," *J. Pers. Soc. Psychol.*, vol. 111, no. 5, pp. 686–705, Nov. 2016, doi: 10.1037/pspi0000066.

[9] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: Design and collection of a multilingual emotional speech database," *Proc. 3rd Int. Conf. Lang. Resour. Eval. LREC 2002 2024–2028*, p. 5.

[10] J. A. Hinojosa *et al.*, "Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions," *Behav. Res. Methods*, vol. 48, no. 1, pp. 272–284, Mar. 2016, doi: 10.3758/s13428-015-0572-5.

[11] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018, doi: 10.1016/j.neucom.2018.05.005.

[12] A. Tharwat, "Classification assessment methods," *Appl. Comput. Inform.*, vol. ahead-of-print, no. ahead-of-print, Aug. 2020, doi: 10.1016/j.aci.2018.08.003.

[13] S.-O. Caballero-Morales, "Recognition of Emotions in Mexican Spanish Speech: An Approach Based on Acoustic Modelling of Emotion-Specific Vowels," *Sci. World J.*, vol. 2013, pp. 1–13, 2013, doi: 10.1155/2013/162093.

[14] H. Pérez-Espinosa, J. Martínez-Miranda, I. Espinosa-Curiel, J. Rodríguez-Jacobo, L. Villaseñor-Pineda, and H. Avila-George, "IESC-Child: An Interactive Emotional Children's Speech Corpus," *Comput. Speech Lang.*, vol. 59, pp. 55–74, Jan. 2020, doi: 10.1016/j.csl.2019.06.006.

[15] K. Hammerschmidt and U. Jürgens, "Acoustical Correlates of Affective Prosody," *J. Voice*, vol. 21, no. 5, pp. 531–540, Sep. 2007, doi: 10.1016/j.jvoice.2006.03.002.

[16] J. A. Hinojosa, E. M. Moreno, and P. Ferré, "Affective neurolinguistics: towards a framework for reconciling language and emotion," *Lang. Cogn. Neurosci.*, vol. 35, no. 7, pp. 813–839, Sep. 2020, doi: 10.1080/23273798.2019.1620957.