Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 368–375
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Turkish Emotion-Voice Database (TurEV-DB)

Salih Fırat Canpolat, Zuhal Ormanoğlu, Deniz Zeyrek
Graduate School of Informatics, Cognitive Science Department, Middle East Technical University (METU),
Ankara, Turkey
salih.canpolat@metu.edu.tr, zuhal.ormanoglu@metu.edu.tr, dezeyrek@metu.edu.tr

Abstract

We introduce the Turkish Emotion-Voice Database (TurEV-DB), which involves a corpus of over 1735 tokens based on 82 words uttered by human subjects in four different emotions (*angry, calm, happy, sad*). The speech data were produced by amateur actors, checked by assessors, recorded, and preprocessed by a denoising procedure. An emotion corpus was constructed on the basis of the finalized recordings. The database involves this corpus together with spectrograms, extracted prosodic features, prosodic data graphs overlaid onto spectrograms, and model activations. Three machine learning experiments were run using a convolutional neural network (CNN) and a support vector machine (SVM) model. Frequency-filtering was applied on the validation set, resulting in speech signals in three frequency ranges: 0-8000 Hz, 0-5000 Hz, and 500-8000 Hz. Frequency filtering is motivated by the fact that frequencies up to 8000 Hertz provide adequate information to human beings to detect sounds, frequencies up to 5000 Hz provide the necessary information about speech sounds, and frequencies below 500 Hz lack most of the speech energy. We describe the proposed methodology for constructing the corpus and the database, report the performance of the machine learning models, and for evaluation, compare machine learning results with human judgements.

Keywords: emotion, speech resource, machine learning

1. Introduction

Emotion is an essential part of interpersonal communication as well as human-machine interaction, and such information can be obtained by processing the features of speech (Rozgić et al., 2012). Emotion recognition can be described as predicting the emotional information, category-wise or factor-wise, from the speech signal (Kim, et al., 2013).

As recording technology and computer processing power improved, the study of emotion has become computerized, and the relation between speech and emotion has been studied in this context extensively. To date, emotion corpora of different languages (encompassing both single languages and multiple languages) have been compiled. For example, Berlin Emotional Speech Database (Burkhardt et al., 2005) consists of a single language, German, and has 800 tokens. INTERFACE (Hozjan et al., 2002) on the other hand, includes English, Slovenian, Spanish, and French, and has 175 to 190 tokens for each language. Each corpus offers different properties regarding emotion.

In the context of Turkish, the link between emotion and voice has recently become quite popular. Among the studies that have been conducted, we can mention (Fidan, 2007) who focused on emotion in prosody. Regarding Turkish emotion resources, Meral et al. (2003) have been one of the first to form a database of four emotions in the Boğaziçi University Emotional database (BUEMDB, cf. Kaya & Karpov (2018)). The resource was based on recorded sentences, which were then analyzed to reveal the F_0 contours of the investigated emotions. Turkish Emotional Speech Database (TurES) (Oflazoglu & Yildirim 2013) offers a database of 5305 Turkish utterances recorded from Turkish movies and tagged with seven emotion categories. The EmoSTAR Database was developed by Parlak, et al. (2014) and includes over 300 spoken samples gathered from the TV or the internet. Korkmaz & Atasoy (2016) used Mel-Frequency Cepstral Coefficients (MFCC) to investigate the emotional content of the speech signal based on the EmoSTAR. Furthermore, Bakir (2017) and Bakır & Yuzkat (2018) gathered a voice-corpus of approximately 3740 Turkish voice samples of words and clauses of differing lengths collected from 25 males and 25 females. The authors developed hybrid machine learning models based on the voice samples. A

Name	Size	# of Actors	Source	Emotions	Emotion Axes	Statistical Features	OpenSmile Package
TurES	5305 tokens	582	Turkish Movies	Afraid, Angry, Neutral, Other, Sad, Surprised	Valence, Arousal, Dominance	Yes	emo_large
EmoSTAR	393 tokens	393	TV and Internet	Angry, Neutral, Happy, Sad	None	Yes	emo_base
BUEMODB	484 files	11	Actors*	Angry, Happy, Sad, Unemotional	None	No	No
Voice Corpus	3740 samples	50	Not known	Afraid, Angry, Sad, Happy, Neutral	None	Yes	No

Table 1: Comparison of overviewed Turkish emotion-voice databases (*non-professional actors).

comparative analysis of the overviewed Turkish emotion voice databases is presented in Table 1.

However, the field still needs more studies. Particularly needed are new databases. The overarching goal of the present study is to introduce Turkish Emotion-Voice Database, or TurEV-DB for short. TurEV is a database of Turkish emotion produced by amateur actors recruited for the sole purpose of the present study.

The specific aims of this study are three-fold: (a) to contribute to the field by compiling a spoken Turkish corpus of words reflecting four emotion categories (*angry, happy, calm, and sad*). These four emotion states were selected as they are generally used in emotion recognition studies such as Eun, et al. (2007), (b) in three experiments, to classify the emotions using two machine learning models; namely, the CNN and the SVM models, and compare the results with human judgements, (c) to form a database that includes the corpus and various peripherals such as spectrograms, continuous and spectral features of the words (e.g. F_0), MFCC, and intermediate activations of the CNN model. In the present study, we describe the corpus and the components of the TurEV database, including how we implemented the two machine learning models. Finally, we compare the performance of the models to each other and to human judgements. Thus, the corpus is evaluated, and the database is validated.

The rest of the paper is arranged as follows: Section 2 describes the data collection process and the database building procedure; Section 3. overviews the machine learning models; Section 4. illustrates the evaluation of the data by human judges; Section 0. presents and discusses the findings, and finally, Section 6. provides a summary and concludes the study.

2. Proposed Methodology

The proposed methodology for TurEV-DB consists of two main parts: The creation of the corpus, and the construction of the database. The creation of the corpus is presented in Section 2.1, and the construction of the database is presented in Section 2.2.

2.1 The Corpus

The creation of the corpus started with selection of words by the authors and proceeded with recording different emotional states by the amateur actors. The steps taken in the corpus creation procedure is presented in Figure 1 and described in the rest of the present section.

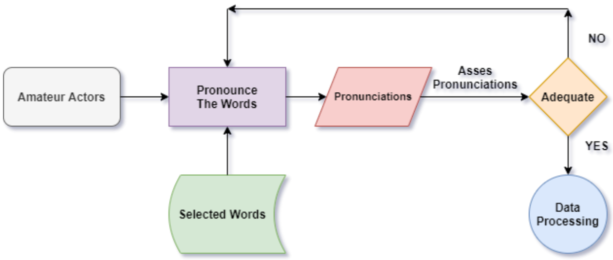

Figure 1: Corpus creation procedure.

2.1.1 Word Selection Procedure

In the literature, there are three types of widely used approaches for constructing an emotion-voice corpus. In the first approach, the desired emotions are induced in the participants by an experimental task; then, emotional speech is collected. The second approach uses the voice of trained or amateur actors who are instructed to speak imitating a specific emotion. The third and the least commonly used method is collecting emotion-induced speech from private recordings and call centers. Since this approach raises ethical concerns and privacy issues, it is not preferred so often. The first method is more natural but less reliable compared to the second approach. We adopted the second approach for our study.

We set out to select the words keeping in mind their phonological profile and selected 82 words from *Türkçenin Ses Dizgesi* (Ergenç & Bekar Uzun, 2017). The selected words involve a range of phonemes, both vowels and consonants, used in various positions in the word. For example, /r/ appears in word-initial and word-final positions as in *rıhtım* /rɯhtɯm/ ‘dock’ and *demir* /demir/ ‘iron’, respectively. Such phonological variation allows the machine learning models to be tested for different conditions and increases the robustness of the models. Moreover, these words have already been investigated by Ergenç et al. (2017) in their neutral (calm) emotional state, and their F_0 values, density graphs, and spectrograms up to 5000 Hz are available. In short, this set of words provided the perfect data and the baseline of our experiments. Table 2 presents the number of tokens in the corpus.

Emotion	**Actor ID**						**Total**
	7895	**1984**	**1234**	**1358**	**1157**	**6783**	
Angry	82	82	82	82	77	82	487
Calm	80	82	82	82	0	82	408
Happy	29	82	82	82	0	82	357
Sad	82	82	82	82	73	82	483
Total	273	328	328	328	150	328	1735

Table 2: Total number of tokens in four emotional states performed by amateur actors.

2.1.2 Amateur Actors

We chose to work with individuals with no previous acting experience to elicit natural speech. From now on, these individuals will be referred to as amateur actors. Using the guidelines, the recruited amateur actors uttered the words and recorded them.

Six amateur actors[1] (three females, three males) produced the selected words. The age of amateur actors was balanced gender-wise to be able to represent the male and the female vocal properties. The ages ranged from the early twenties to mid-thirties. This range enabled us to obtain information from a variety of age groups.

2.1.3 Recording Procedure

The actors followed a set of guidelines prepared by the authors, and conducted their recording sessions on their

[1] The amateur actors were volunteering graduate or undergraduate students in Cognitive Science and Psychology departments.

own. They pronounced the words as if they were in each emotional state. They used the Sound Recorder application from Sony Mobile Productions and recorded the words in mono, WAV format, and in 44100-Hz sampling rate. The actors then listened to their own recordings and presented the recordings to an assessor, a third-party individual of their choosing. The pronunciations that lacked any emotional tone according to the actor or the assessor were re-recorded by the amateur actors. In this regard, each pronounced word is considered a token in TurEV-DB. Moreover, each token has relevant data points extracted from it or produced by it, as explained in the following sections.

2.1.4 Preprocessing

The data processing procedure consisted of two stages, namely the denoising stage and the frequency filtering stage for the experiments. In the denoising stage, the tokens were first denoised, and then they were trimmed. The resulting tokens had uniform silence before and after the word and had relatively low noise.

2.1.4.1 Denoising and Preprocessing

We removed the noise from the recordings and preprocessed them using Audacity 2.3.0[2] through the following steps:

1. The recording of each word was loaded into Audacity.
2. The signal view was changed from waveform into spectrogram.
3. The part that contained no voice was selected, and a noise profile was generated. The generated profile was used as a base profile in order to remove the noise from the recording.
4. The recording was trimmed leaving 150 milliseconds of silence at the beginning; the end and the rest of the recording were removed.
5. The denoised and preprocessed recording was exported in 32-bit Float Pulse Code Modulation (PCM)[3] WAV format.

In this stage, the statistical noise removal technique was used (Audacity Team, n.d.). The procedure could be summarized as the identification of the noise floor for each frequency by analyzing at least 2048 samples, and then removing this noise floor from the entirety of the recording.

2.1.4.2 Frequency Filtering

The data were first split into two, namely, the training and the validation sets. The validation data set consisted of 20% of the tokens and was used in the CNN model, the generation of the activation maps, and in human validation. Frequency filtering was applied to the validation tokens, resulting in speech signals in three frequency ranges: 0-8000 Hz, 0-5000 Hz, and 500-8000 Hz. Machine learning experiments were run on the frequency-filtered data.

Frequencies within 0-8000 Hz contain most of the acoustic features of the human voice. Although a young adult can hear more than 16000 Hz, such high frequencies do not contain useful features for human speech. Also, the 0-8000 Hz range is needed for the optimum amount of information transformation. According to the Nyquist-Shannon theorem (Yadav, 2009; Yao, 2014), in order to describe a signal, it should be sampled at least twice of the Nyquist frequency. In other words, a sine wave with 10 Hz frequency cannot be described by any sampling rate below 20 Hz. However, at 20 Hz, this description yields minimum amount of data. With 22050 Hz Nyquist frequency, we used approximately 1/3 of this maximum value, which is 8000 Hz. This strategy allowed us to be able to work both with a data-rich environment and relatively high frequencies. Thus, for the first experiment, the frequencies over 8000 Hz were trimmed out, and only the frequencies between 0 and 8000 Hz were kept.

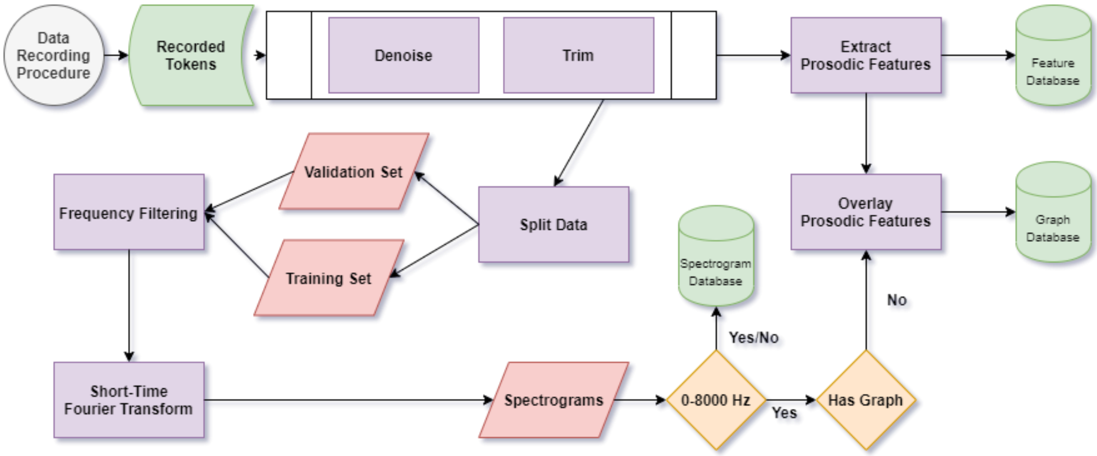

Figure 2: Construction of the database.

Given that frequencies up to 5000 Hz actually contain the most useful features for humans, for the second experiment, the data were filtered trimming frequencies over 5000 Hz and the experiment was run on the 0-5000 Hz range. For the third experiment, the frequencies under 500 Hz were trimmed. This was motivated by the fact that most of the speech energy, as well as F_0 is cut out when the frequencies under this level are blocked. From an empirical perspective, comparing the predictive ability of the CNN model with the humans would be interesting in the third experiment, as it would reveal whether the model and

[2] https://www.audacityteam.org/audacity-2-3-0-released/

[3] https://docs.microsoft.com/en-us/windowshardware/drivers/audio/pcm-stream-data-format

humans perform similarly in the absence of low frequencies.

2.2 Construction of the Database

Construction of the database included the process of feature extraction, spectrogram generation, and overlaying the features onto the spectrograms. The construction of the database also overlapped with the creation of the corpus at the data preprocessing stage. A summary of the database construction procedure is presented in Figure *2*.

2.2.1 Feature Extraction

Spectral features were extracted from the preprocessed recordings using OpenSmile[4], a freely available application Eyben et al. (2015). Low-level descriptors such as MFCCs and $F_{0,}$ and the statistical functions applied to them were extracted, amounting to a total of 1582 extracted features. (These features were also used in the SVM model.) Moreover, time-variant acoustic features were also extracted. The procedure is elaborated in Section 2.2.2.

2.3 Acoustic Information

external data, e.g. by means of tokens generated from the same or a different word pool, and even data from different languages.

We extracted further acoustic features using OpenSmile. F_0, voicing probability, and loudness features were extracted with a 50 milliseconds window size and 10 milliseconds step size. These features were projected onto spectrograms for further inspection. The extracted features and spectrograms with acoustic information were added to TurEV-DB. This addition will allow future researchers to inspect the tokens with easy-to-read graphics instead of using hard to read data formats such as CSV. An example of such a spectrogram with projected acoustic features is presented in Figure *3*.

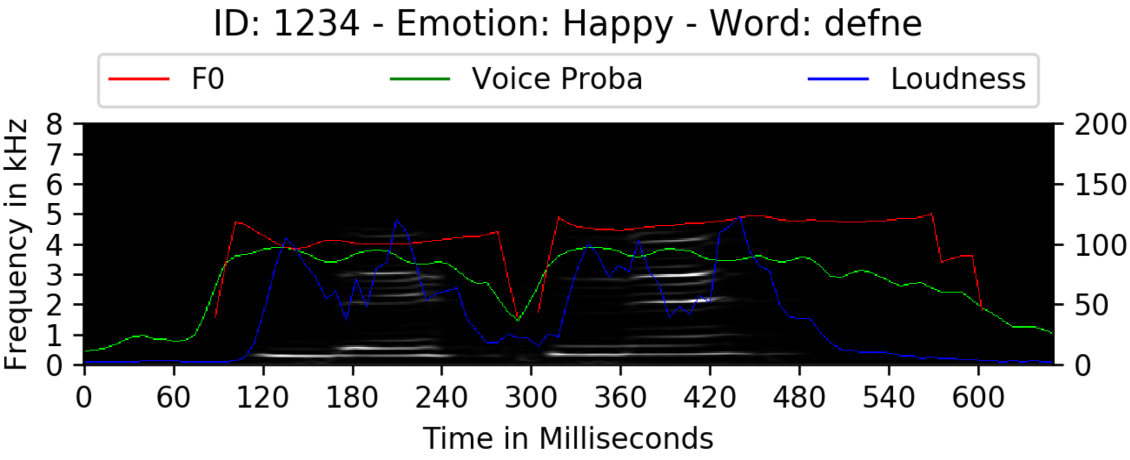

Figure 3: The spectrogram of the token defne ‘happy' (uttered in a happy emotional state) and the acoustic features projected on it.

2.2.2 Spectrogram Generation

A spectrogram generation procedure specifically tailored for convolutional neural networks was used in this stage. The script was applied on all of the corpus tokens including the frequency-filtered tokens of the validation set. This procedure consisted of a short-time Fourier transformation (STFT) with Hamming window (Podder, et al., 2014). A 2205- sample long window size, which amounts to 50 milliseconds with 95% overlap was used in the STFT algorithm. In order to determine the optimal spectrogram output, we used different amounts of overlaps and different amounts of sample sizes in our pilot analyses. Each recording was saved as 746x495 pixel PNG image with 32-bit pixel depth in the grayscale colourmap. The spectrograms, as well as the script used for this procedure, were included in the TurEV-DB so that any set of new data could be processed using the script. These components allow the current and future models to be tested using

3. Machine Learning Models

Three experiments were run on frequency-filtered data using a CNN and SVM model. A CNN and an SVM model were constructed since they can be initialized and used on the go with the data included in TurEV-DB. We chose the SVM and the CNN model types because they represent different machine learning paradigms and offer different advantages.

3.1 The SVM Model

The SVM model was initialized with the RBF kernel and one versus rest (OVR) decision function. 10 features were extracted from the initial 1582 features (Section 2.2.1) using principal component analysis (PCA). 80% of the data was selected for training, 10% was used for validation, and 10% for testing. The results of the test set are presented in Table 4.

[4] https://www.audeering.com/opensmile/

3.2 The CNN Model

The CNN model was built using a custom set of layers, as shown in Figure 4. It exploits stacked layers of convolution operators without max-pooling in order to produce high-resolution activation maps. The CNN model accepts spectrograms of the words and outputs emotion category probabilities and the activation maps. As we explain briefly below, the activation maps are the result of Grad-CAM operations which allow the model to output its internal state regarding specific data and conditions, e.g. for the pronunciation of *rıhtım* ‘dock’ in the emotional state *happy*.

The results of the CNN model are also provided in Table 4 and indicate that with 76% accuracy, the CNN model is the highest performer compared to the SVM model as well as the human judgments for the data in the frequency range between 0-8000 Hz. It performs still well, though with a slightly reduced power (72%) in the frequencies between 0-5000 Hz. However, its accuracy plummets to 49% when it is subjected to spectrograms in the frequency range of 500-8000 Hz. The CNN model required 90 minutes for the 9 epochs of training.

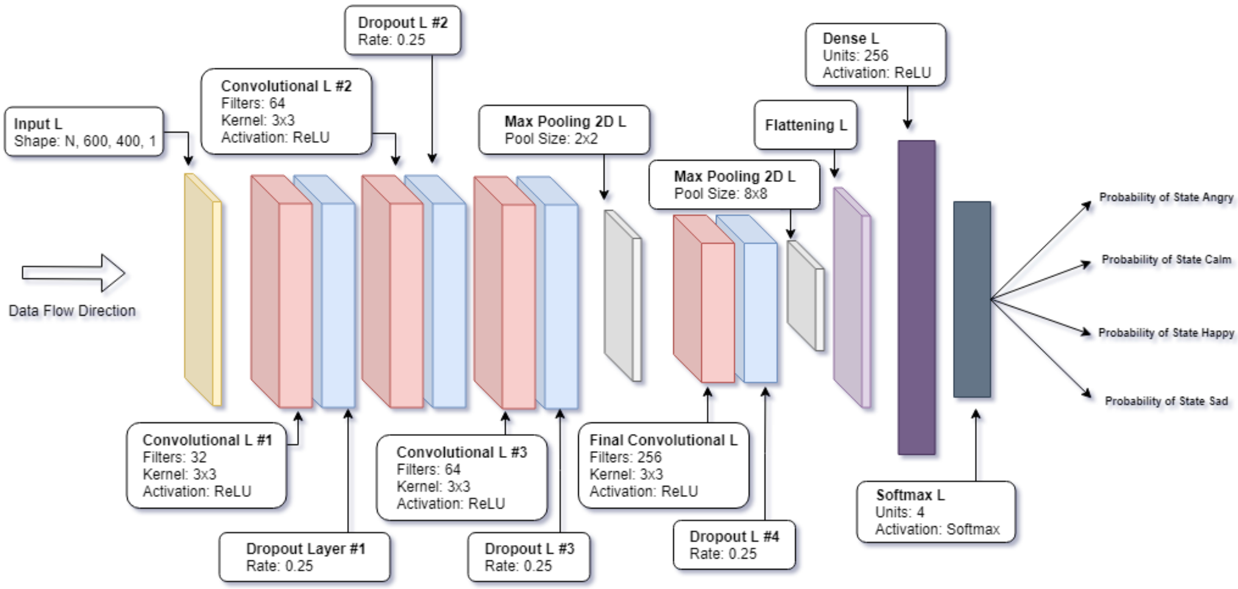

Figure 4: The architecture of the CNN model.

Fold #	Accuracy	
	Maximum	9th Iteration
1	0.70	0.65
2	0.74	0.71
3	0.75	0.72
4	0.77	0.77
5	0.75	0.72
6	0.79	0.78

Table 3: Accuracy values for the cross-validation study for the CNN model.

3.3 The Gradient-Weighted Class Activation Mapping (Grad-CAM) Model

The Grad-CAM model (Selvaraju et al., 2017) is a sub-part of the CNN model. It uses the final convolutional layer of the CNN model and projects intermediate activations according to an emotional state in this layer. The projections of intermediate activations are then visualized. These visualizations, which are also called heat maps, will allow users to inspect the areas that contribute most to the decision-making process of the CNN model.

3.4 Validation

The SVM model follows the standard 80% training, 10% validation, and 10% testing splits. The CNN model, on the other hand, uses 6-fold cross-validation. A new CNN model with the same architecture was generated and subjected to training and validation processes for each fold. The final accuracy of the CNN model was found to be within the bounds of the results obtained with the cross-validation study. Therefore, the model can be considered valid within the bounds of TurEV-DB. The results of the cross-validation study are presented in Table 3.

4. Experimental Results and Analysis

4.1 Data Evaluation by Human Judges

For data evaluation, three volunteering human judges were recruited. Each of the three judges was presented with the data of one of the experiments.[5] They were simply asked to listen and group the tokens into one of the four emotional states. They were presented the tokens with pseudo-names and not given any information about their emotional content. The matching decisions were considered correct and mismatching decisions were considered incorrect. The

[5] The judges were volunteering graduate students with a Bachelor’s degree in psychology.

results of human evaluation are presented in Table 4 and show that they are quite stable across experiments.

4.2 Comparative Analysis of The Results of The CNN Model and The Human Judges

To compare the predictions of the CNN model and the human judgements, contingency tables were created for the frequency ranges investigated in the experiments and presented in Table 5, Table 6, and Table 7. These tables indicate that through different experiments involving different frequency ranges, the model and the judges converge on the categorization of two emotions, namely *happy* and *angry*. The emotion category *happy* is agreed by the CNN model and the judge with a ratio of 68%, 76.47%, and 71.05% in the different frequency ranges, respectively. Similarly, the emotion category *angry* is agreed by the CNN model and the judge with a ratio of 61.90%, 47.06%, and 75.23% in the different frequency ranges, respectively. Regarding the categorization of *angry*, in the frequency range 0-5000 Hz, the power of the model and the judges is lower than the other frequency ranges.

	Frequency Bands (in Hertz)			
	0-8000	**0-5000**	**500-8000**	
Judges	0.65	0.64	0.64	**Accuracy Rating**
CNN Model	0.76	0.72	0.49	
SVM Model	0.61	-	-	

Table 4: Classification accuracy of the models compared to human judgments in different frequency ranges.

CNN Model	**The Human Judge**			
	Angry	**Calm**	**Happy**	**Sad**
Angry	*61.90*	10.69	10.00	6.35
Calm	19.05	*38.17*	14.00	19.05
Happy	14.29	16.79	*68.00*	14.29
Sad	4.76	34.35	8.00	*60.32*

Table 5: Frequency distribution of emotion categorization by CNN and the human judges in the 0-8000 Hertz band.

CNN Model	**The Human Judge**			
	Angry	**Calm**	**Happy**	**Sad**
Angry	*47.06*	14.73	3.92	1.02
Calm	14.71	*37.21*	13.73	24.49
Happy	32.35	22.48	*76.47*	10.20
Sad	5.88	25.58	5.88	*64.29*

Table 6: Frequency distribution of emotion categorization by CNN and the human judges in the 0-5000 Hertz band.

CNN Model	**The Human Judge**			
	Angry	**Calm**	**Happy**	**Sad**
Angry	*75.23*	42.98	21.05	43.18
Calm	1.83	*9.65*	2.63	3.41
Happy	16.51	35.96	*71.05*	29.55
Sad	6.42	11.40	5.26	*23.86*

Table 7: Frequency distribution of emotion categorization by CNN and the human judges in the 500-8000 Hertz band.

To reveal the CNN model's and the judges' classification performance, precision, recall and F1 scores were calculated separately using the database labels as the key. The results are presented in Table 8, Table 9 and, Table 10. The metric results presented in Table 8 indicate that the CNN model succeeded in classifying all four emotions when it used the widest frequency band (0-8000 Hz).

According to Table 9, when the CNN model is fed with a signal lacking frequencies above 5000 Hz, it tends to classify *angry* less accurately with a recall of 0.51. However, according to Table 10, when the CNN model is fed with a signal lacking the frequencies below 500 Hz, it outright classifies *angry* with a recall value of 0.84. A similar performance is displayed by the judges for *angry* with a recall of 0.79. In this frequency range, for *calm*, the CNN model yielded a very low recall score of 0.11, and the judges a low recall score of 0.57. The category *sad* also has very low recall score of 0.31, compared to 0.67 of the judges (Table 10).

	The Emotion Categories							
	Angry		**Calm**		**Happy**		**Sad**	
Metrics	M	J	M	J	M	J	M	J
Precision	0.84	0.76	0.69	0.43	0.71	0.80	0.79	0.83
Recall	0.76	0.80	0.74	0.68	0.79	0.56	0.75	0.54
F1	0.80	0.77	0.71	0.53	0.75	0.66	0.77	0.65

Table 8: Metrics for CNN's and the judges' emotion classification for the 0-8000 Hz frequency band.

	The Emotion Categories							
	Angry		**Calm**		**Happy**		**Sad**	
Metrics	M	J	M	J	M	J	M	J
Precision	0.93	0.90	0.67	0.38	0.60	0.84	0.77	0.69
Recall	0.51	0.63	0.73	0.60	0.85	0.61	0.82	0.71
F1	0.66	0.74	0.70	0.46	0.71	0.70	0.80	0.70

Table 9: Metrics for CNN's and the Judges' emotion classification for the 0-5000 Hz frequency band.

	The Emotion Categories							
	Angry		**Calm**		**Happy**		**Sad**	
Metrics	M	J	M	J	M	J	M	J
Precision	0.46	0.71	0.53	0.41	0.41	0.92	0.70	0.74
Recall	0.84	0.79	0.11	0.57	0.64	0.49	0.31	0.67
F1	0.60	0.74	0.18	0.48	0.50	0.64	0.43	0.70

Table 10: Metrics for CNN's and the Judges' emotion classification for the 500-8000 Hz frequency band.

In summary:

- In 0-5000 Hz and 0-8000 Hz frequency ranges, the CNN model has a higher score than the judges as well as the SVM model. In the 500-8000 Hz frequency range, the CNN model underperforms, whereas the judges do not have any performance loss (see Table 4).
- According to Tables 5-7, while the emotion states *happy, angry* have relatively high classification rates both by the CNN model and the judges, the emotion states *calm* and *sad* have low classification scores. In contrast to *sad*, *calm*, *angry* and *happy* are high energy emotions with high-frequency outputs.
- Regarding the recall metrics of the human judges and the CNN model, Table 8 shows that the CNN model uses the widest frequency range of 0-8000 Hz to easily differentiate between the emotions *angry* and *happy* well as *calm* and *sad.* Table 10 indicates that the CNN model uses the 500-8000 Hz frequency band to recall the emotion category *angry* with a higher success than the emotion category *happy*.

5. Discussion

Beyond forming a corpus of 1735 tokens in four different emotion states and a database that includes low-level descriptors and acoustic features, the current study has several conclusions and implications. Regarding the machine learning models and the advantages they offer, the CNN model is computationally expensive; therefore, it is slow. On the other hand, the SVM model is computationally cheap, and consequently fast. Both models rely on extracted low-level descriptors and their statistical derivations as features but an SVM model cannot perform in a sample where low frequencies are missing or when there is highly noisy data; in SVM, such features result in the extraction of only noise or blank features. On the other hand, the CNN model can perform over such data. In this regard, the CNN model is more versatile for our purposes. Performance-wise, in frequency ranges 0-8000 Hz, the CNN model outperforms both the human judges and the SVM model, where the SVM model performs the poorest. Another advantage of the CNN model is its ability to produce intermediate activations (heat maps). These intermediate activations can be used both for computational studies and manual inspection. The CNN model is computationally expensive, yet it offers an advantage over SVM by outperforming SVM in the range 0-8000 Hz. Moreover, the CNN model succeeds in the frequency range 0-5000 Hz, where the SVM cannot produce any result. Given that the 0-5000 Hz frequency range involves the frequency band that sufficiently represents human speech, the performance of the CNN model in this frequency range is a good sign that it can be of use in future research.

The low prediction results of CNN in the 500-8000 Hz frequency range may have some implications on human beings' perception of emotion in words. In this frequency range, the CNN model's performance, but not the human judges, substantially decreased. This result may suggest that human beings are not affected by the absence of the missing acoustic properties below 500 Hz in classifying the words into four emotion states, while the CNN model is. In fact, in this frequency range, the CNN performed well in classifying *angry* and to a lesser degree, *happy* (with a recall of 0.64). The overall low performance of CNN in this range is affected by the low recall scores for *calm* (0.11) and *sad* (0.31) (Table 10). Given that *calm* and *sad* are low energy emotional states, we can speculate that when frequencies below 500 Hz are blocked, the CNN model seems affected negatively because, among other acoustic features, speech energy is indeed lost below 500 Hz. Further research is called for to establish what factors precisely led to this result.

6. Conclusion

In this study, we mainly described the development of the Turkish Emotion-Voice Database (TurEV-DB), a database that integrates a core corpus of emotion-laden word pronunciations with peripheries. In this study, we mainly described the development of the Turkish Emotion-Voice Database (TurEV-DB) which resulted a voice corpus and a database package consisting of different features. The corpus component includes 1738 tokens generated from 4 emotions (angry, calm, happy, and sad) and 82 words by 6 amateur actors. The dataset package carries a wide range of components.

- 1582 statistical features extracted using OpenSmile and IS10_paralig configuration.[6]
- 3-time variant prosodic features extracted using OpenSmile and prosodyAcf configuration.
- 1738 spectrograms generated using STFT.
- 1738 spectrograms overplayed with prosodic features.
- An SVM machine learning model.
- A CNN deep learning model.
- 349 heat maps derived from intermediate activations of the CNN model.

The present study was limited by the small number of actors (and therefore tokens) as well as the small number of judges. In the future, we plan to increase the number of tokens and the judges, and develop better evaluation models to increase the quality of the tokens. We also intend to enrich the database by including more recordings from new amateur actors. With the completion of the future work TurEV-DB will be made open to the public.

[6] https://www.audeering.com/download/opensmile-book-latest/

7. Acknowledgements

We would like to thank Şiyar Morsünbül, Kübra Yılmaz, Oğuzcan Ülgen, Cansu Aşıcı, Ayşe Nur Balli, Dilşah Suretli, Gamze Gücenmez Üngörmüş, Mustafa Özaydın, Enis Dönmez, Emre Erçin, Baki Cağdaş¸ Aydın, Reyyan Baş Ümit Murat, Dilara Uslu, Müge Çelikel, Buşra Gizem Sönmez, and İrem Yıldız for their assistance in data collection, and Cengiz Acartürk and İclal Ergenç for useful comments on an earlier version of the study. We thank three anonymous reviewers for their insightful comments. Any errors are our own.

8. Bibliographical References

Audacity Team. (n.d.). Audacity Manual. Retrieved February 4, 2019, from https://manual.audacityteam.org/man/noise_reduction.html

Bakir, C. (2017). Speech recognition system for Turkish language with hybrid method. *Global Journal of Computer Sciences: Theory and Research, 7*(1), 48. https://doi.org/10.18844/gjcs.v7i1.2699

Bakır, Ç., & Yuzkat, M. (2018). Speech Emotion Classification and Recognition with different methods for Turkish Language. *Balkan Journal of Electrical and Computer Engineering*, *6*(2), 54–60. https://doi.org/10.17694/bajece.419557

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., Berlin, T. U., Berlin, H. U. (2005). A Database of German Emotional Speech. In *Proc. of INTERSPEECH 2005*.

Ergenç, İ., & Bekar Uzun, İ. P. (2017). *Türkçenin Ses Dizgesi* (1st ed.). Ankara: Seçkin Yayıncılık.

Eun, H. K., Kyung, H. H., Soo, H. K., & Yoon, K. K. (2007). Speech emotion recognition using eigen-FFT in clean and noisy environments. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (pp. 689–694). https://doi.org/10.1109/ROMAN.2007.4415174

Eyben, F., Wöllmer, M., Schuller, B. B., Weninger, F., Wollmer, M., & Schuller, B. B. (2015). OPENSMILE: open-Source Media Interpretation by Large feature-space Extraction. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*. https://doi.org/10.1145/1873951.1874246

Fidan, D. (2007). Türkçe ezgi örüntüsünde duygudurum ve sözedim görünümü [Emotion and speechacts in Turkish intonation pattern] (Unpublished Ph.D. Thesis). Ankara University.

Hozjan, V., Zdravko, K., Asuncion, M., Antonio, B., Albino, N., Moreno, Z., … Nogueiras, A. (2002). Interface Databases: Design and Collection of a Multilingual Emotional Speech Database. In *LREC* (pp. 2019–2023). Las Palmas de Gran Canaria, Spain.

Kaya, H., & Karpov, A. A. (2018). Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing, 275,* 1028–1034. https://doi.org/10.1016/j.neucom.2017.09.049

Kim, Y., Lee, H., & Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2013.6638346

Korkmaz, O. E., & Atasoy, A. (2016). Emotion recognition from speech signal using mel-frequency cepstral coefficients. In *ELECO 2015 - 9th International Conference on Electrical and Electronics Engineering* (pp. 1254–1257). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ELECO.2015.7394435

Meral, H. M., Ekenek, H. ., & Özsoz, A. (2003). Analysis of emotion in Turkish. In *XVIIth National Conference on Turkish Linguistics*.

Oflazoglu, C., & Yildirim, S. (2013). Recognizing emotion from Turkish speech using acoustic features. *Eurasip Journal on Audio, Speech, and Music Processing*. https://doi.org/10.1186/1687-4722-2013-26

Parlak, C., Diri, B., & Gürgen, F. (2014). A Cross-Corpus Experiment in Speech Emotion Recognition Yildiz Technical University , Turkey. In *2nd Workshop on Speech, Language and Audio in Multimedia (SLAM 2014)* (pp. 11–12). Retrieved from https://www.isca-speech.org/archive/slam_2014/papers/slm4_058.pdf

Podder, P., Zaman Khan, T., Haque Khan, M., & Muktadir Rahman, M. (2014). Comparative Performance Analysis of Hamming, Hanning and Blackman Window. *International Journal of Computer Applications*. https://doi.org/10.5120/16891-6927

Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Vembu, A. N., & Prasad, R. (2012). Emotion recognition using acoustic and lexical features. In *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012* (Vol. 1, pp. 366–369). Retrieved from http://www.isca-speech.org/archive

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2017.74

Yadav, A. (2009). Nyquist-Shannon Sampling Theorem. In *Digital Communication*.

Yao, Y.-C. (2014). Nyquist Frequency. In *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat03517