

# Beyond Words: Interjection Classification for Improved Human-Computer Interaction (2023)

Y. Goren, A. Apartsin, Y. Cohen, Y. Aperstein

**Abstract—** In the realm of human-computer interaction, fostering a natural dialogue between humans and machines is paramount. A key, often overlooked, component of this dialogue is the use of interjections such as “mmm” and “hmm”. Despite their frequent use to express agreement, hesitation, or requests for information, these interjections are typically dismissed as “non-words” by Automatic Speech Recognition (ASR) engines. Addressing this gap, we introduce a novel task dedicated to interjection classification, a first in the field to our knowledge. This task is challenging due to the short duration of interjection signals and significant inter and intra-speaker variability. In this work, we present a dataset of interjection signals collected specifically for interjection classification. We employ this dataset to train and evaluate a baseline deep learning model. To enhance performance, we augment training dataset using techniques such as tempo and pitch transformation, which significantly improve classification accuracy, making models more robust.

**Index Terms—** Human-computer interaction, Automatic Speech Recognition, Multiclass Classification, Audio Data Augmentation, Robustness

## I. INTRODUCTION

PEOPLE tend to choose an efficient form of verbal communication by leveraging common understanding of context between two parties. Interjections are one of the major parts of speech frequently used to convey meaning in a specific context. Many smart systems include a speech recognition system that enables natural dialogue between man and machine [1]. Among well-known examples are virtual personal assistants, including Apple's Siri artificial intelligence system [2], or Amazon's Alexa [3]. A Siri or Alexa user can, with natural speech queries, obtain information and perform various actions in several domains (e.g., checking weather or stock prices, ordering pizza, etc.).

In recent years, a communications revolution has expanded from a rich human-machine interaction to a full-fledged person-machine communication [1], [2], [3], [4]. This revolution requires a high-quality interface to support successful Human Computer Interactions (HCI) [4]. Adding interjection recognition capabilities to voice assistants will improve human computer interactions and increase usage, with spontaneous conversation in human-machine interfaces. The improvement

can be expressed by interpreting the meaning of hard-to-understand speech, such as a heavy accent, or a sentence in which not all words are clear [5]. By understanding prominent interjections, we can understand the whole sentence. For example, when a human says “Oy, ...” in a conversation, he wants to express an unexpected situation, something unfortunate, or perhaps fright.

Interjections often express spontaneous feelings, a key feature that can be harnessed for Speech Emotion Recognition (SER) [6]. While humans adeptly recognize emotional cues in speech, this remains a research frontier for machines. Enabling machines to comprehend and respond to emotions can enhance services like call centers, particularly in emergency situations. This emotional intelligence in machines is crucial for fostering more human-like interactions.

Voice interjections can be seen as a form of “voice touch” signal, akin to haptic touch interfaces, potentially simplifying interactions with voice assistants and other voice-enabled devices. Interjections can be mapped to specific actions or responses, much like shortcuts in a user interface. For instance, positive interjections like “aha” could affirm a process's continuation, while “uh-oh” could signal a need for repetition. This allows for the replacement of lengthy commands with simple interjections, customizing the system to individual user preferences and needs. For example, a user could choose an interjection to represent a frequently used command, streamlining their interaction with the system.

It is important to emphasize that our aim is not to detect specific interjection phrases within speech or to build a system that recognizes all interjections. Instead, our research presents a benchmark system for interjection classification, providing a solid foundation for the quick and simple addition of new interjections. Our motivation is to enhance user interface technologies, enabling the creation of more natural and habitable human-machine interfaces and dialogues.

The collection and the preparation of the training data is a major challenge for this project. The available speech datasets are focused on word level [7], phoneme level [8], or event level tasks [9], while interjections lie somewhere in-between. Therefore, we collected our own unique baseline dataset by recording a relatively large set of interjections and negative

Y. Goren is with the Intelligent Systems Department, Afeka Tel-Aviv Academic College of Engineering, 69988. E-mail: yanivgoren@gmail.com.

A. Apartsin is with the Intelligent Systems Department, Afeka Tel-Aviv Academic College of Engineering, 69988. E-mail: apartsin@gmail.com

Y. Cohen is with the Industrial Engineering Department, Afeka Tel-Aviv Academic College of Engineering, 69988. E-mail: yuvalc@afeka.ac.il

Y. Aperstein is with the Intelligent Systems Department, Afeka Tel-Aviv Academic College of Engineering, 69988. E-mail: apersteiny@afeka.ac.il

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

examples from some speakers.

In this work, we propose a neural network model for interjection classification. Since deep learning models require a large amount of training data, we enriched the dataset by augmenting the original data through the application of various artificial distortions [10]. The data augmentation includes the addition of background noise [11] and pitch and tempo modifications [12].

In this paper, we present results of several different interjection classification baseline experiments, where relative improvement obtained by using the proposed data augmentation methods. By incorporating these augmentation methods into the training process, we enhanced the network's ability to handle variations, thereby improving its capacity to generalize and perform effectively on new, unseen data.

The rest of this paper is organized as follows: We begin with a review of prior work, focusing on speech recognition, augmentation methods in this field, and research related to interjections. Subsequently, we provide an in-depth explanation of this research contribution. This is followed by an overview of the dataset creation process and a description of our data augmentation tool. We then introduce the interjection classification models and delve into the specifics of our experimental setup, which is designed to evaluate the effectiveness of our approach in various scenarios. Finally, we discuss the results, summarize the proposed method, and suggest potential avenues for future research.

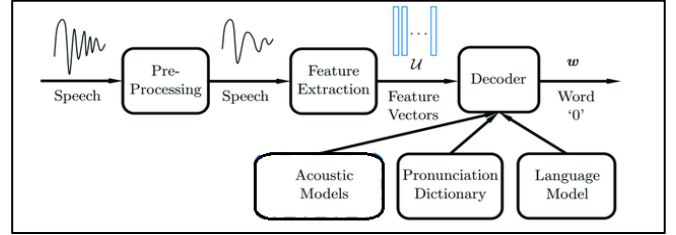
## II. PREVIOUS WORK

One of the most common tasks associated with processing speech and voice signals is Automatic Speech Recognition (ASR) [13]. ASR is the task of translating audio signals into text. The main challenge is to overcome the non-stationarity of the speech signal and the large variations in its spatio-temporal representation. As illustrated in Figure 1, the typical ASR system usually includes several steps. The pre-processing step, where an analog speech signal is transformed into a digital signal, includes speech/non-speech segmentation and filtering. The feature extraction step deals with the transformation of the incoming digital waveform into a vector representation of desirable speech features that emphasize linguistic information. Usually, a speech signal is broken into short (usually around 20-30 ms) segments (frames), overlapped every 10 ms, during which the signal is assumed to be stationary [14], [15]. During the last few decades, several techniques have been developed for feature extraction from speech signals. These approaches include Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear prediction (PLP), Relative Spectral (RASTA), and Linear Predictive Coding (LPC) [16]. However, MFCCs are probably used most commonly used [17].

MFCCs extract spectral features by defining an analysis window (around 25 ms) and divide the speech signal into different time frames by shifting the window with a 10 ms shifting stride. Then Fast Fourier Transformation (FFT) is calculated for each frame to obtain the frequency features, and the logarithmic Mel-Scaled filter bank is applied to its

power spectrum estimate. MFCCs calculate the Discrete Cosine Transformation (DCT) of the log energies in the corresponding frequency bands to obtain an  $m$ -dimensional coefficients vector. The measured power spectrum envelope in each frame correlates to the shape of the vocal tract, providing an appropriate representation of the sound or phoneme being produced. This procedure results in feature vectors that can be arranged in a  $[n \times m]$  matrix, where  $m$  is the number of coefficients and  $n$  is the number of frames.

At the heart of an ASR system is the decoder. During this phase, feature vectors are decoded into linguistic units that make up speech. The decoding relies on acoustic and language models [18] to recover the most probable utterance by modeling the conditional probability of the  $n$ th word, using the  $(n-1)$  earlier words. Linguistic and pronunciation dictionaries are often used to improve the decoding performance. An acoustic model [18] is a fundamental part of the ASR system, where the connection between the acoustic information (e.g., feature vectors) and phonetics is established. The acoustic models are frequently implemented using various methods that include the Hidden Markov Model (HMM) [19] and support vector machines (SVM) [20]. HMM is the most commonly used acoustic model for speech recognition in many practical applications.



**Fig. 1.** Automatic speech recognition system diagram.

ASR is a broad topic that includes subtopics related to interjection recognition. One of those subtopics is Key-word Spotting (KWS) [21]. The need to provide users with a fully hands-free experience for situations like driving resulted in the development of a system that listens continuously for specific keywords to initiate voice input. Keyword Spotting aims at detecting predefined keywords in an audio stream. A commonly used technique for keyword spotting is the Keyword/Filler Hidden Markov Model (HMM). One disadvantage with this technique is that it can be computationally expensive, and the model is trained separately for each keyword [22].

Keyword spotting is sometimes performed using a pattern matching approach where the input is compared with a few prerecorded commands. Since the same word might be articulated with a different speed, the Dynamic Time Warping (DTW) algorithm is used to align two sequences in an optimal way [23]. Recent neural network models show a significant improvement over the HMM approach. The recurrent neural networks (RNN) used in the Deep KWS model. [22] shows good performance while keeping a reduced runtime computation and smaller memory footprint. A Convolutional

Neural Network (CNN) architecture described by [24] shows improvements over FNN in a variety of small and large vocabulary tasks. The described CNN architecture generalizes more easily to different speaking styles compared to a fully connected FNN architecture.

Another sub-topic related to interjection recognition is Audio Event Detection (AED) [9], which is considered a common task of processing speech and voice signals. An audio event is a specific type of sound, such as footsteps, running water, exhaust fan noise, screams, ocean waves breaking, or music. Many sound clips contain multiple acoustic events that overlap on the time axis. AED is a task in which a relatively long (several seconds to tens of seconds) sound clip including multiple acoustic events serves as input, and the output is acoustic event labels and their time stamps (start and end times). The process of AED is created by extracting acoustic features using MFCC in general, and then constructing classification models using HMM [19], SVM [20] or more recently models such as CNN [24] or RNN [25]. Because acoustic events can often have a temporal overlap, the most difficult problem in AED is how to detect active durations of acoustic events. For instance, Mulimani et al. [26] proposed acoustic event detection based on a convolutional neural network, which calculates a posterior for the existence of acoustic events time frame by time frame.

Generally, deep learning requires a large amount of labeled training data to enable accurate speech recognition. To the best of our knowledge, such a large dataset does not exist for interjections. Therefore, data augmentation is proposed, where the speech data are artificially augmented by applying different types of distortions in a way that does not change the label. Stoian et al. [13] and Singh et al. [27] proposed data augmentation for low resource speech recognition tasks. The performance of an FNN speech recognition model depends on how well the training data matches the testing data. This can be done by increasing the generalization of a model beyond the data provided to it, and by attempting to create similarity between the training data and the representative characteristics that are seen in real data, such as voice variability of different speakers, or different back-ground noises.

There are many options for data augmentation. Some of them are applied in the feature level of a neural network, and some are applied directly in the raw audio level. Vocal tract length perturbation (VTLP) [28] is a popular method for doing feature level data augmentation in speech and has shown gains on the TIMIT (a collection of phonemically transcribed American English speech) phoneme recognition task. SpecAugment [29] is another feature level data augmentation method for speech recognition that operates on the log Mel spectrogram of the input audio. In the raw audio level, intuitive and practical transformations, such as Dynamic Range Compression (DRC), pitch-shifting, time stretching, and background noise combination apply audio effects to the original training audio files [30].

Despite the prevalence of interjections in human speech patterns, interjections in literature are mentioned mainly in the context of emotion where researchers seek to understand the

nature of interjections and interjectional meaning [31],[32]. There is a difference between formal speech and conversational speech. Conversational speech is more spontaneous and efficient and does not require special training. Székely [33] describes fundamental properties of spontaneous speech that present problems for spoken language applications, such as lack of punctuation or the inability to “hear” a speaker’s emotion through speech. As stated before, SER is still an ongoing subject of research and the main traditional techniques for SER are based on feature extraction and selection to identify various emotions [6].

No prior work, to our knowledge, has explored inserting interjections in speech recognition systems. Cohn et. al. [34] described an experiment that systematically manipulated the Amazon Alexa TTS by adding some emotional-cognitive expressions, like interjections, while also examining the influence of interjection duration and pitch levels. The conclusion was that those manipulations improved user’s ratings of their conversation across thousands of real user interactions. Likewise, an emotional speech recognition system that classified “fear emotion” for systems like emergency call centers is described by [35]. This system used Support Vector Machines (SVM) with an interjections feature, that can be classified as a type of incredulity observed in spontaneous speech when a speaker gets hyperemotional. It is important to clarify that this system cannot recognize different interjections but rather knows how to classify calls into two classes: calls that include fear and calls that do not include fear.

### III. OUR CONTRIBUTION

In this work, we propose a novel approach to the task of interjection classification, employing a fully connected feedforward neural network (FNN). The task presents unique challenges, given the nature of interjections as short, context-free sound fragments that exhibit significant variability within and between speakers.

To our knowledge, no existing dataset is specifically tailored to the study of interjections. As such, we have taken the initiative to create a new dataset comprised of interjection phrases from various speakers. This dataset serves as a valuable resource for research in this area.

Furthermore, we introduce the use of audio data augmentation as a strategy to mitigate the limitations posed by data scarcity. We investigate the impact of different augmentation techniques on the performance of our proposed architecture. By training the network on augmented data, we aim to enhance the network’s robustness to deformations and improve its generalizability to unseen data in both quiet and high-noise environments.

### IV. METHODOLOGY

Our aim is to create a baseline machine learning model for interjection classification. This model will process audio input and assign labels, distinguishing between interjections and non-interjection words. To train this model, we created a ‘clean’ dataset from various speakers. Initially, we use data from a single speaker to focus on interjection recognition. The dataset

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

is later expanded with more speakers and augmented with artificial modifications to pitch, tempo, and added background noise.

#### A. Construction of Original Dataset

Our dataset consists of 'clean', unsynthesized audio samples spanning five distinct categories: four dedicated to different interjections and one to non-interjection words. The latter category encapsulates words randomly selected from a book. The interjection categories were specifically chosen due to their language-independence and their ability to convey inherent semantic meaning, irrespective of conversational context. Table I presents the selected interjections.

TABLE I  
FOUR SELECTED INTERJECTIONS THAT RECORDED  
INTO SEPARATE SOUND CLIPS

interjection	alternate/similar	translation/meaning
nah		"No" - Informal no
mmm	mhm, uh-hu	"Yes" - Agreement, acknowledgement
ahah	Aha, ahh	"I understand" - Understanding, Confirmation
oy	oy vey	"Oh no..." - Mainly Jewish: Used to express self-pity, or expression of unexpected situation

To serve the purposes of this research, we recorded five speakers, comprising two females and three males, each contributing a certain number of audio samples for every category. The column "Number of Audio Samples" refers to a specific word, e.g., Speaker A has 850 audio samples per word and a total number of 4,250 audio samples. The speaker profiles and the count of audio samples contributed by each speaker for every category are detailed in Table II. A profile denoted as a triple (X, Y, Z) signifies X as the gender (M for male, F for female), Y as the age, and Z as the speaker's mother tongue (H for Hebrew, S for Spanish).

TABLE II  
SPEAKER PROFILES AND NUMBER OF AUDIO  
SAMPLES RECORDED FOR EACH SPEAKER AND  
WORD

SPEAKER	NUMBER OF AUDIO SAMPLES	PROFILE
A	~850	M, 44, H
B	~550	F, 42 H
C	~300	M, 16, H
D	~300	F, 81, S
E	~300	M, 50, H

The sound clips were recorded using Audiotk, which is a Voice Activity Detection tool that enable recording and saving each sound clip as a separate wav file with 16000 Hz sampled rate.

#### B. Data Preprocessing

All audio files in our dataset were standardized to a uniform length. This was done to ensure a consistent number of frames across all files during the feature extraction phase, resulting in fixed-length vectors for our classification model. We chose a length of 1.55 seconds, which accommodates longer interjection terms. We also set a minimum recording length of 0.45 seconds to exclude short, unrelated recordings such as background noises, and a maximum length of 1.55 seconds to account for instances where Audiotk did not detect silence between recordings. Any recordings falling outside this length range were excluded from the dataset.

To achieve a fixed length, silence was added to the beginning and end of each sound. This ensured that the actual sound portion of each recording was unevenly distributed across the fixed length. For instance, one recording might begin with 0.2 seconds of silence and end with 0.74 seconds of silence, while another of the same label might start with 0.61 seconds of silence and conclude with 0.3 seconds of silence. This approach prevents the model from learning the exact onset of the word's sound. Following this process, we generated five folders, corresponding to four interjections and one negative example, each containing approximately 800 fixed-length recordings.

To extract the useful features from the audio file, Librosa library [34] has been used. This library is a python package for music and audio analysis that provides several methods to retrieve information from sound clips. The methods from Librosa we have been used to extract various features are: MFCC, Melspectrogram, Chorma-stft, Spectral contrast and Tonnetz. The result of that process is a matrix with a row for each audio file, and a column for each mean feature value (193 column).

#### C. Data Augmentation System (DAS)

To extend the collected dataset, a data augmentation system (DAS) that generates synthetic samples was developed. We used Pysox (Ashar et. al., 2020), a Python library that provides a simple interface between Python and Sox, a popular command line tool for sound processing that can apply various effects to audio files. We tried four different sound effects on our original unsynthesized

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

(“clean data”) audio samples. Each effect was applied directly to the sound file prior to converting it into the input vector representation for the neural network. The data augmentation effects are described be-low:

- *Tempo*: Change the audio playback speed but not its pitch. This function gets the parameter ‘factor’ (factor > 1 speed up the audio signal, factor < 1 slow down the audio signal). Duration of the sound file is changing.

- *Pitch*: Change the audio pitch (but not tempo). The sensation of a frequency is commonly referred to as the pitch of a sound. A high pitch sound corresponds to a high frequency sound wave and a low pitch sound corresponds to a low frequency sound wave. One octave (the interval between one musical pitch and another) is divided into 12 semitones (tones) of 100 cents each. Typically, cents are used to express small intervals. 1200 cents equal to 1 octave. This function gets the parameter ‘n-semitones’ (The positive or negative number of semitones to shift). Duration of the sound file is not changing.

- *Background noise (BGN)*: Mixes the original “clean” audio with another recording containing background sounds from several acoustic scenes. Each sample is mixed with nine acoustic scenes: baby gibberish, ambulance, crowd laughing, football crowd, mall, passing bus, rain, street traffic, and TV. Each mix  $S_{mix}$  was generated using (1)

$$S_{mix} = (W_{orig} \times S_{orig}) + (W_{BGN} \times S_{BGN}) \quad (1)$$

where  $S_{orig}$  is the original audio sample and  $S_{bkg}$  is the signal of the background scene.  $W_{orig}$  and  $W_{BGN}$  are weighting volume parameters determined such that  $W_{orig} + W_{BGN} = 1$ . For each  $S_{orig}$  audio, two or three  $S_{mix}$  audio files with different weights were generated, depending on the dataset.

DAS can create a large number of synthetic samples from original “clean” samples. The number of synthetic samples is determined by the configuration of the methods to use (tempo, pitch, and background noise are the three supported methods now, but any other method can be easily added), and by configuration of each method value. For example, if the configuration for the pitch method contains four different values for the semi-tone parameter, then the system will generate four different augmented audio samples with the desired pitch level.

A significant advantage in DAS is the fact that different effects can be combined for each recording. This allows us to control the size and the diversity of the dataset we want to create; e.g., if the desired methods are tempo (with factors 0.9 and 1.1) and pitch (with semitones 2 and -2), an additional eight audio samples will be created from one original audio (two audios with only tempo effect, two with only pitch and four that are result of combining tempo and pitch parameters together).

In addition, DAS can generate white noise in the background

of each original audio file. This functionality was added because white noise is produced by combining sounds of all different frequencies together and can be used to mask other sounds, such as back-ground noises.

#### D. Network architecture and Evaluation metric

In our research, we employ a fully connected feedforward neural network (FNN) to classify single-word audio files into one of five categories, four of which are interjections and one non-interjection. This architecture, implemented via TensorFlow, is tailored to the multiclass nature of our task. The softmax activation function in the output layer generates a probability distribution vector of potential outcomes, facilitating the classification process.

To measure the accuracy of our model, we utilize the F1 score, which balances precision and recall. This metric is particularly suitable for our task as it accounts for both false positives and false negatives. Therefore, our model is not only designed to accurately identify interjections but also to minimize misclassifications, a critical aspect for enhancing human-computer interaction.

### V. EXPERIMENTAL SETUP

Our experimental setup is designed to evaluate the effectiveness of our approach in various scenarios. We first present augmented datasets using the Data Augmentation System (DAS). We then present two main evaluation scenarios devised to examine the impact of data augmentation on the effectiveness of interjection classification in both quiet and high-noise environments.

#### A. Construction of Augmented Datasets using DAS

It is important to choose the augmentation parameters such that the semantic validity of the label is maintained. To create our augmented datasets, we chose several parameters: for tempo, a factor in the range of 0.86-1.14; for pitch, a semitone in the range of (-2.4)-2.4; and for background noises,  $W_{orig}$  in the range of 0.83-0.93 and  $W_{background}$  in the range of 0.07-0.17. In our research, we developed augmented datasets by applying tempo, pitch, and background noise effects, both individually and in various combinations. The resulting augmented sets are presented in Table III.

For each speaker, 120 different audio samples were used from our original dataset (600 files altogether from four interjection folders and one folder of negative examples), whose overall length is about eight minutes, to generate seven augmentation sets with a total of 180,000 audio samples, whose overall length was about 78 hours. For all speakers, the system generated more than 300 hours of augmentation data from 32 minutes of original recordings.

TABLE III  
DESCRIPTION OF THE AUGMENTED DATASETS

	ORIGINAL SAMPLES PER CLASS	METHODS USED	GENERATED SAMPLES PER CLASS
1	120	Pitch	2520
2	120	Tempo	2520
3	100	BGN	2900
4	60	Tempo + Pitch	4860
5	40	Pitch + BGN	6880
6	40	Tempo + BGN	6880
7	10	Tempo + Pitch + BGN	9320

### B. Evaluation Scenarios

Our study is structured around two primary scenarios. The first scenario is designed to assess how data augmentation can enhance the performance of interjection classification for speakers not previously encountered by the model. We initiate this scenario by training the classifier using clean data from two speakers (one male and one female). The model is then tested on clean data from two different speakers (also one male and one female). The data from the remaining subject is utilized as a validation set for optimizing hyperparameters. Subsequently, the classifier is trained separately on each of our seven augmented datasets and retested on the clean data from the unseen speakers.

In the second scenario, we explore the impact of data augmentation on classification performance in real-world, high-noise environments. We posit that a system trained with augmented data can effectively handle real external sound sources. The initial phase of this scenario involves training the classifier using clean data from a single speaker. The model is then tested on data derived from audio signals recorded from the same speaker in a real-world noisy environment, which includes background noises such as a television playing, children's sounds, a baby crying, and an air conditioner running. In the subsequent phase, the classifier is trained using our artificially augmented data sets from the same speaker. The model's accuracy is then tested against the same natural background data set used in the first phase of this scenario.

The models were trained using the Adam optimizer, with a learning rate of 0.009 and 3 hidden layers. These parameters were selected after extensive experimentation with various alternatives. For instance, we tried training epochs ranging from 150 to 300, adjusted the learning rate starting from 0.1 and reducing it for better convergence, experimented with different activation functions, number of hidden layers and tested different optimizer algorithms like classic SGD and Adam.

## VI. RESULTS AND DISCUSSION

Two experiments were conducted in our study. In both the accuracy assessment was measured by comparing the baseline (original dataset) accuracy with the accuracy level of each of the proposed augmentation methods.

The results of the first experiment are presented in Table IV. The model trained with samples of two subjects (female speaker A and male speaker B) and tested separately on previously unseen speakers (male unseen speaker C and female unseen speaker D). The data from subject E was utilized as a validation set for optimizing hyperparameters. The percentages within the parentheses present the accuracy improvement compared the baseline clean dataset accuracy. We can realize that each of the augmented methods significantly improves the accuracy relative to the baseline accuracy. Each of the augmented sets were helpful in this scenario, and the highest classification accuracy improvement for each unseen speaker was achieved by the dataset combined with all three augmentation methods, where an increase of 63.6% was obtained for speaker C and 42.9% was obtained for speaker D, which confirm our claim that the combination of the augmentation methods gives better results.

TABLE IV  
RESULTS OF THE FIRST SCENARIO

TRAINING DATASET (SPEAKERS A+B)	ACCURACY (SPEAKER C)	ACCURACY (SPEAKER D)
Original clean data	0.286	0.41
Pitch	0.391 (36.7%)	0.44 (7.3%)
Tempo	0.405 (41.6%)	0.477 (16.3%)
BGN	0.418 (46.2%)	0.415 (1.2%)
Tempo + Pitch	0.456 (59.4%)	0.438 (6.8%)
Pitch + BGN	0.454 (58.7%)	0.484 (18%)
Tempo + BGN	0.393 (37.4%)	0.529 (29%)
Tempo + Pitch + BGN	0.468 (63.6%)	0.586 (42.9%)

The results of the second experiment are presented in Table V, where the impact of augmentation in natural high noise environment was tested. After training the classifier with clean data of one speaker and testing it with natural background noises, we trained it with the augmented datasets we created for the same speaker. We average accuracy over all subjects. Highest classification accuracy improvement achieved by combining tempo and artificial background noises, where in several of the testing the accuracy was beyond 0.7. However, in contrast to the strong impact artificial background noises, and the slight positive effect of the tempo, the impact of the pitch was negative. A better future experiment will be to try different semitones values for this augmentation method.

TABLE V  
RESULTS OF THE SECOND SCENARIO.

TRAINING DATASET	ACCURACY
Original clean data	0.45
Pitch	0.344 (-23.6%)
Tempo	0.476 (5.8%)
BGN	0.63 (40%)
Tempo + Pitch	0.424 (-5.8%)
Pitch + BGN	0.569 (26.4%)
Tempo + BGN	0.667 (48.2%)
Tempo + Pitch + BGN	0.567 (26%)

In both scenarios the model seems to be more robust when using the background noise method. In most cases, and more precisely 6 from 9 cases, the combination between background noises and other methods improved the accuracy of the classifier in relation to use only the background synthesis. It implies that using some different settings for each effect can improve the results even more than the results we obtained in the scenarios described.

## VII. CONCLUSION

In this research, we have ventured into the largely unexplored territory of interjection classification in human-computer interaction. Recognizing the importance of these often-dismissed "non-words" in fostering natural dialogue, we have introduced a novel task dedicated to their classification. This task, however, is not without its challenges, given the short duration of interjection signals and the significant variability within and between speakers.

To tackle these challenges, we created a unique dataset of interjection phrases from various speakers and have developed a model for interjections classification. The constructed dataset, the first of its kind, serves as a valuable resource for further research in this area.

Recognizing the limitations posed by data scarcity, we have also introduced the use of audio data augmentation. By training our network on augmented data, we have enhanced its robustness to deformations and improved its generalizability to unseen data. This approach has proven particularly effective in both quiet and natural high-noise environments.

Our findings underscore the potential of our approach to enhance human-computer interaction by incorporating interjections, paving the way for more natural and intuitive dialogues that truly reflect the nuances of human communication.

## REFERENCES

[1] M. B. Hoy, "Alexa, Siri, Cortana, and more: an introduction to voice assistants", *Medical reference services quarterly*, vol.37, no. 1, pp. 81-88,2018 , doi: 10.1080/02763869.2018.1404391

[2] A. Berdasco, G. López, I. Diaz, L. Quesada, , L. A. Guerrero, "User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana". In: *Multi-disciplinary Digital Publishing Institute Proceedings*, vol. 31, no. 1, Art no. 51,

2019, doi: <https://doi.org/10.3390/proceedings2019031051>

[3] I. Lopatovska, K. Rink, I. Knight, K. Raines, K. Cosenza, H. Williams, A. Martinez, "Talk to me: Exploring user interactions with the Amazon Alexa". *Journal of Librarianship and Information Science*, vol. 51, no. 4, pp. 984-997, 2019, doi: <https://doi.org/10.1177/0961000618759414>

[4] L.Clark, P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, & R. B. Cowan, "The State of Speech in HCI: Trends, Themes and Challenges". *Interacting with Computers*, vol. 31, no. 4, pp. 349-371,2019 . <https://doi.org/10.1093/iwc/iwz016>

[5] S. K. Gouda, S. Kanetkar, D. Harrison, & M. K. Warmuth, "Speech Recognition: Keyword Spotting Through Image Recognition". *arXiv preprint arXiv:1803.03759*.2018 . <https://doi.org/10.48550/arXiv.1803.03759>

[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar & T. Alhussain, "Speech emotion recognition using deep learning techniques: A review". *IEEE Access*, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[7] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition". *arXiv preprint, arXiv:1804.03209*.2018 . <https://doi.org/10.48550/arXiv.1804.03209>

[8] A. Almadhor, R. Irfan, J. Gao, N. Saleem, H. T. Rauf, & S. Kadry, "E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition". *Expert Systems with Applications*, vol. 222, Art. no. 119797,2023 , <https://doi.org/10.1016/j.eswa.2023.119797>

[9] K. Imoto, "Introduction to acoustic event and scene analysis". *Acoustical Science and Technology*, vol. 39, no. 3, pp. 182-188, 2018. <https://doi.org/10.1250/ast.39.182>

[10] A. Beghdadi, M. Mallem, L. Beji, "Benchmarking performance of object detection under image distortions in an uncontrolled environment". In *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, pp. 2071-2075, October 2022. doi: 10.1109/ICIP46576.2022.9897643

[11] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, P. Gamble, "Voices obscured in complex environmental settings (voices) corpus". *arXiv preprint arXiv:1804.05053*,2018 . <https://doi.org/10.48550/arXiv.1804.05053>

[12] K. R. Kulkarni, & S. R. R. Naik. "A Review of Music Analysis Techniques", *International Research Journal of Engineering and Technology*, vol. 5, no. 3, pp. 4102-4108, March. 2018. <https://www.academia.edu/download/56794754/IRJET-V5I3962.pdf>

[13] M. C. Stoian, S. Bansal and S. Goldwater, "Analyzing ASR Pretraining for Low-Resource Speech-to-Text Translation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, Barcelona, Spain, pp. 7909-7913, May 2020. doi: 10.1109/ICASSP40776.2020.9053847

[14] Y. Hu, P. N. Samarasinghe, S. Gannot and T. D. Abhayapala, "Semi-Supervised Multiple Source Localization Using Relative Harmonic Coefficients Under Noisy and Reverberant



- Environments," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3108-3123, 2020, doi: 10.1109/TASLP.2020.3037521
- [15] C. Ick and B. McFee, "Sound Event Detection in Urban Audio with Single and Multi-Rate Pcen," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, Toronto, Canada, pp. 880-884, June 2021, doi: 10.1109/ICASSP39728.2021.9414697
- [16] P. Gadekar, M. H. Kaldane, D. Pawar, O. Jadhav, & A. Patil, "Analysis of speech recognition techniques", *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 5. No. 2. pp. 1129-1132, 2019. <https://www.academia.edu/download/59152470/V5I2-158420190506-30272-1e4jqre.pdf>
- [17] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi & A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech", *Biomedical Signal Processing and Control*, vol. 71, part A, Art. no. 103107, 2022, doi: 10.1016/j.bspc.2021.103107,
- [18] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, & M. Almojel, "Automatic speech recognition: Systematic literature review", *IEEE Access*, vol. 9, pp. 131858-131876, 2021, doi: 10.1109/ACCESS.2021.3112535.
- [19] B. Mor, S. Garhwal, & A. Kumar, "A systematic review of hidden Markov models and their applications". *Archives of computational methods in engineering*, vol.28, pp. 1429-1448, 2021. <https://doi.org/10.1007/s11831-020-09422-4>
- [20] A. Bhavan, P. Chauhan, & R. R. Shah, "Bagged support vector machines for emotion recognition from speech". *Knowledge-Based Systems*, vol. 184, Art. no. 104886, 2019. <https://doi.org/10.1016/j.knosys.2019.104886>.
- [21] I. López-Espejo, Z. H. Tan, J. H. Hansen, & J. Jensen, "Deep spoken key-word spotting: An overview", *IEEE Access*, vol. 10, pp. 4169-4199, 2021. doi: 10.1109/ACCESS.2021.3139508.
- [22] K. H., Chang, P. H. H. Huang-Yu, Y. Jin, & T. C. Wang, "Audio adversarial examples generation with recurrent neural networks". In *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Beijing, China, pp. 488-493, January 2020. doi: 10.1109/ASP-DAC47756.2020.9045597.
- [23] M. Yadav, & A. Alam, "Dynamic time warping (DTW) algorithm in speech: a review", *Int. J. Res. Electron. Comput. Eng.*, vol. 6., no. 1, pp. 524-528, 2018.
- [24] J. Li, "Recent advances in end-to-end automatic speech recognition. APSIPA", *Transactions on Signal and Information Processing*, vol. 11, no. 1, pp.1-61, 2022. .
- [25] S. P. Yadav, S. Zaidi, A. Mishra, & V. Yadav, "Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN)". *Archives of Computational Methods in Engineering*, vol. 29, no. 3, pp. 1753-1770, 2022. <https://doi.org/10.1007/s11831-021-09647-x>
- [26] M. Mulimani, A. B. Kademani, & S. G. Koolagudi, "A deep neural network-driven feature learning method for poly-phonic acoustic event detection from real-life recordings". In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 291-295, May 2020, doi: 10.1109/ICASSP40776.2020.9054208.
- [27] S. Singh, R. Wang, & F. Hou, "Improved Meta Learning for Low Resource Speech Recognition". In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, Singapore, pp. 4798-4802, May 2022. doi: 10.1109/ICASSP43922.2022.9746899.
- [28] C. Kim, M. Shin, A. Garg, & D. Gowda, "Improved Vocal Tract Length Perturbation for a State-of-the-Art End-to-End Speech Recognition System". In *Interspeech*, pp. 739-743, September 2019. doi: 10.21437/Interspeech.2019-3227
- [29] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, & Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition". *arXiv preprint arXiv:1904.08779*, 2019. <https://doi.org/10.48550/arXiv.1904.08779>
- [30] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279-283, March 2017. doi: 10.1109/LSP.2017.2657381
- [31] R. A. Ulfatovna, "Interjections oh! wow! gosh! ah! god! as a means of expressing emotions of personages in the novel 'can you keep a secret'." by S. Kinsel-la. *ACADEMICIA: An International Multidisciplinary Research Journal*, vol. 11, no. 1, pp. 865-870, 2021. doi: 10.5958/2249-7137.2021.00163.4,
- [32] K. B. Nabila, & O. T. Ena, "The analysis of interjection types and emotional state used in "Sam & Cat" sitcom". *Rainbow Journal of Literature, Linguistics and Culture Studies*, vol. 11, no. 2, 9-15, 2022. <https://doi.org/10.15294/rainbow.v11i2.55333>
- [33] É. Székely, G. E. Henter, J. Beskow, & J. Gustafson, "Spontaneous Conversational Speech Synthesis from Found Data". In *Interspeech* pp. 4435-4439, September 2019. doi: 10.21437/Interspeech.2019-2836
- [34] M. Cohn, C. Y. Chen, & Z. Yu, "A Large-Scale User Study of an Alexa Prize Chatbot: Effect of TTS Dynamism on Perceived Quality of Social Dialog". In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 293-306, September 2019. doi: 10.18653/v1/W19-5935
- [35] S. A. Yoon, G. Son & S. Kwon, "Fear emotion classification in speech by acoustic and behavioral cues". *Multimedia Tools and Applications*, vol. 78, no. 2, pp. 2345-2366, 2019. <https://doi.org/10.1007/s11042-018-6329-2>
- [36] D. A. Barreda, J. E. Lainez., D. Sharma, & P. Naylor, "U.S. Patent No. 9,997,172". Washington, DC: U.S. Patent and Trademark Office, 2018. <https://patents.google.com/patent/US9997172B2/en>



**Yaniv Goren** received B.Sc. from the Department of Computer Science at Hadassah Academic College Jerusalem (2005), M.Sc. in Intelligent Systems from The Afeka Academic College of Engineering (2020). He has over 15 years of industry development of computer systems in the field of Voice Recognition. He currently serves as a VoiceAI team leader at AudioCodes company.

**Alexander Apartsin** received B.Sc. from the Technion (1995), M.Sc. from Weizmann Institute of Science (2004), Ph.D. in Computer Science from Tel-Aviv University (2012) and did his post-doctoral research at the Weizmann Institute of Science. He has over 25 years of industry research experience in the fields of Machine Learning, Signal and Image Processing and Speech Analysis. He is currently supervising students at Afeka Academic College of Engineering in Tel Aviv

**Yuval Cohen** received B.Sc. from Ben-Gurion University (1989), M.Sc. from the Technion-IIT (1992) and Ph. D from the University of Pittsburgh (1998) in Industrial Engineering. From 1998 to 2001 Yuval Cohen served as a Senior Operations Planner at FedEx Ground (USA) responsible for the five-year plan for the 24 hubs and 400 terminals. From 2002 to 2014 he served as the Head of the Industrial Engineering program at the Open University of Israel. From 2014 Yuval Cohen is a senior Professor (Ph. D) at the Industrial Engineering department of the Tel-Aviv Afeka Academic College of Engineering. Prof. Cohen is a fellow of the Institute of Industrial Engineers (IIE), and a Full Member of the Institute for Operations Research and Management Sciences (INFORMS).

**Yehudit Aperstein** received B.Sc. from Haifa University(1995), M.Sc. from the Technion(1998), Ph.D. in Mathematics from Weizmann Institute of Science(2005) and did her post-doctoral research at Bar Ilan and Tel-Aviv University. She is currently the Head of Intelligent Systems Graduate Program at Afeka Academic College of Engineering in Tel Aviv. Her current research interests include robust AI for intelligent systems and applications of deep learning in intelligent systems.