To Invest or Not to Invest: Using Vocal Behavior to Predict Decisions of Investors in an Entrepreneurial Context

Goossens, Ilona ; Jung, Merel M. ; Liebregts, Werner; Onal Ertugrul, Itir

Publication date:
2022

Document Version
Peer reviewed version

To Invest or Not to Invest: Using Vocal Behavior to Predict Decisions of Investors in an Entrepreneurial Context

Ilona Goossens[1], Merel M. Jung[1], Werner Liebregts[2], and Itir Onal Ertugrul[3]

[1] Department of Cognitive Science and Artificial Intelligence, Tilburg University, The Netherlands {i.goossens, m.m.jung}@tilburguniversity.edu
[2] Jheronimus Academy of Data Science, The Netherlands
w.j.liebregts@tilburguniversity.edu
[3] Department of Information and Computing Sciences, Utrecht University, The Netherlands
i.onalertugrul@uu.nl

Abstract. Entrepreneurial pitch competitions have become increasingly popular in the start-up culture to attract prospective investors. As the ultimate funding decision often follows from some form of social interaction, it is important to understand how the decision-making process of investors is influenced by behavioral cues. In this work, we examine whether vocal features are associated with the ultimate funding decision of investors by utilizing deep learning methods. We used videos of individuals in an entrepreneurial pitch competition as input to predict whether investors will invest in the startup or not. We proposed models that combine deep audio features and Handcrafted audio Features (HaF) and feed them into two types of Recurrent Neural Networks (RNN), namely Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). We also trained the RNNs with only deep features to assess whether HaF provide additional information to the models. Our results show that it is promising to use vocal behavior of pitchers to predict whether investors will invest in their business idea. Different types of RNNs yielded similar performance, yet the addition of HaF improved the performance.

Keywords: vocal behavior, entrepreneurial decision making, deep learning, VGGish, LSTM, GRU

1 Introduction

Entrepreneurial decision-making is at the core of successfully operating within the business sector [37]. It includes all decisions made by entrepreneurs themselves and decisions made by others which have an immediate impact on the entrepreneur. Due to the complex and dynamic environment (e.g., high uncertainty, ambiguity, time pressure, high risks, etc.) that entrepreneurs and associates (e.g., innovators, investors, inventors) are subject to as well as the frequency with which decisions on entrepreneurial tasks and activities have to

be made, the entrepreneurial decision-making process often relies on flexible decision-making principles. Previous works [25], [18], [2], [13], [39] suggest that when social interactions are involved during the decision-making process, behavioral cues (e.g., physical appearance characteristics, posture and gestures, face and eye movement, and vocal behavior) strongly affect the ultimate decision. Even though many entrepreneurial decisions are made with little to no social interactions, some decisions are heavily based on human-to-human interaction. This is especially true for decisions related to the provision of financial resources by investors in the start-up business environment as entrepreneurial pitch competitions (i.e., events where entrepreneurs convey their start-up business idea to prospective investors) are a common approach to attract financial support. Since these decisions are associated with long-term start-up outcomes, understanding how and to what extent behavioral cues expressed during social interactions influence the decision-making process of investors could benefit entrepreneurs as they could apply this knowledge to increase the effectiveness of their presentation style which, in turn, could lead to an increase in funding [5], [25], [32]. In general, enhancing our understanding of the decision-making process is valuable as decisions have a direct effect on important outcomes for businesses, organizations, institutions, individuals, and societies. Knowledge on how to improve those outcomes would benefit all of these stakeholders [29].

Research on decision-making in the entrepreneurial context is predominantly derived from psychological, sociological, and economic literature. In contrast, research on using machine learning approaches for automated analysis of human behavior to understand the entrepreneurial decision-making process is limited. Previous work focused on applying conventional machine learning methods such as k-Nearest Neighbors (kNN) and support vector machines (SVM) to predict investment based on the visual features including facial actions [28], eye gaze [3], and facial mimicry [19]. Vocal behavior has not been explored in automatically predicting the decisions of investors. Given the superior performance of deep learning-based approaches in several audio classification tasks [17] and the significance of vocal behavior in decision-making process [14], we propose to utilize deep learning methods to model vocal behavior and to predict decisions of investors. This research is conducted on a dataset including video recordings of individuals performing an entrepreneurial pitch about their start-up business idea. They participated in a pitch competition to attract financial resources from potential investors. This study may reveal the importance of vocal characteristics in explaining decisions related to business funding and business growth which have been neglected in research so far [25].

As vocal behavior is derived from speech that has spatiotemporal dynamics, it is crucial to incorporate a deep learning approach with the ability to retain information from previous time points. For that reason, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, which are two types of Recurrent Neural Network (RNN), are incorporated as their performance on sequence-based tasks and capturing long-term dependencies is well established [7], [16]. Besides, LSTM and GRU are both considered to be effec-

tive for recognizing vocal characteristics and classifying audio recordings. The RNN architectures are combined with a Convolutional Neural Network (CNN) where the CNN extracts context-aware deep audio features that are fed into the RNN (e.g., [26], [33], [6]). Combining CNN's capability to learn invariant features with RNN's capability to model temporal dependencies into a single classifier is better known as a Convolutional Recurrent Neural Network (CRNN) [4], and is the current state-of-the-art approach in research on audio classification. In this study, we compare the performance of the models consisting of a CRNN architecture with LSTM or GRU with the proposed models additionally including Handcrafted audio Features (HaF). Introducing HaF into the model can impact the performance as Giannakopoulos et al. [12] reported a significant increase in performance when deep context-aware audio features (i.e., CNN) were combined with HaF. Additionally, Tianyu et al. [40] found that HaF capture complementary information that benefits the RNN.

Considering the literature on entrepreneurial decision-making and the various deep learning approaches across the field of audio classification, this paper explores the fusion of deep context-aware audio features and HaF in combination with the most common architectures for sequence modeling (i.e., LSTM and GRU). Results show that the proposed deep learning models have the ability to detect and recognize vocal patterns which could be used to predict the investor's funding decision. Moreover, this study finds an increase in performance when HaF are introduced into the model, while the impact of the different RNN architectures is negligible as they yielded comparable performances.

2 Related work

Audio processing based on deep learning approaches is an emerging field due to the promising results these methods produced for tasks such as pitch determination [16], audio and sound classification [34], affective speech classification [23], and audio source separation [27]. As data used for audio processing contains prominent sequential signals, a systematic approach that incorporates the ability to capture spatiotemporal dynamics is required [16]. Recurrent Neural Networks (RNN) are suitable for modeling sequential dependencies and nonlinear dynamics within audio data as it is able to retain information from previous allocation due to the recurrent connections within the network that allow for encoding temporal information [16].

The approach of including a RNN for sequential modeling is widely adopted across audio classification field. For example, Chung et al. [7] compares different types of RNN architectures (i.e., LSTM and GRU) to a traditional Deep Neural Networks (DNN) and reveals that the models including the recurrent units outperform the traditional DNN on classification tasks including music and raw speech signal data.

Recent studies propose a Convolutional Recurrent Neural Network (CRNN) where the RNN, which is highly capable learning temporal context and model sequential data, is used in combination with a Convolutional Neural Network

(CNN). The CNN model has proven to be effective in feature learning as it is able to extract shift invariant high level features which could not be modeled with feature engineering.

Traditionally, feature representations are generated from a feature engineering process which requires domain knowledge and relies heavily on researchers' engineering effort for the task at hand. However, research towards feature learning in deep neural networks, which has been of interest lately as it reduces the required expertise and engineering effort, explored the potential of CNN architectures. Comparing the two approaches, a study by Trigeorgis et al. [41] on speech emotion recognition concluded that feature representations created through an end-to-end deep network significantly outperforms the approach of traditional designed features based on signal processing techniques and shallow processing architectures. They argue that deep networks, especially CNNs, have the ability to extract context-aware effective and robust acoustic features which better suit the task at hand, and therefore, improve the performance of the model. Hershey et al. [17] examines the performances of multiple CNN architectures on audio soundtrack classification by proposing analogs of popular CNN networks (e.g., AlexNet, VGG, Inception, and ResNet-50), which have proven to be effective in image classification. With minor modifications to the models, results show that all CNN architectures yield significant performances on audio classification problems. Comparing the performances of the architectures trained on 70M videos with 3.000 labels based on log-mel spectrogram inputs, the best performing architecture incorporates the Inception-V3 model achieving 0.918 Area Under Curve (AUC) while the worst performing architecture employs the AlexNet model achieving 0.894 AUC. The ResNet-50 and VGGish models report 0.916 AUC and 0.911 AUC, respectively. The findings in this study support that convolutional layers in deep neural networks effectively recognize and preserve modulation patterns while omitting small deviations in pitch and timing by training to extract, regardless of the offset frequency and start time, down- and upward moving spectral patterns.

Lim et al. [26] propose a CRNN for rare sound event detection. They incorporate a CNN model for feature learning, which takes log-amplitude mel-spectrogram extracted from the audio as the input feature and analyzes the audio in chunk-level. The extracted features resulting from the CNN model are fed into a two-layer LSTM network for modeling sequential information. The best performing model report an error rate of 0.13 and a F-score of 0.931.

Sainath et al. [33] propose a Convolutional LSTM Deep Neural Network (CLDNN) which is a unified framework that is trained jointly. In other words, they design an architecture that captures information about the input representation at different levels by combining a CNN, LSTM, and DNN. Here, the CNN is used to reduce spectral variation in the input feature, the LSTM network performs the sequential modeling, and the DNN layers are introduces for the vocabulary tasks. Sainath et al. [33] hypothesize an improvement in performance and output predictions when DNN layers are introduced into the CRNN as the mapping between hidden units and outputs is deeper. Their initial pro-

posed method achieved a word error rate of 17.3, which decreased to 17.0 when uniform random weight initialization was introduced to deal with vanishing gradients. Although their proposed model yields better results, the performance of the CRNN model is with a word error rate of 17.6 considered comparable, yet competitive. They conclude that both the CRNN model and the CLDNN model are able to capture information at different resolutions. Additionally, the performance of the CLDNN model was evaluated on a large dataset resulting in an error rate of 13.1. However, this error rate increases to 17.4 when noise is added.

A similar CRNN configuration to that of [33] is adopted by Cakir et al. [4] for a polyphonic sound event classification task. The main difference between the network as proposed in [33] and applied in [4] is the type of RNN architecture incorporated. Cakir et al. [4] includes a GRU network instead of a LSTM network to model the long-term temporal context in the audio. Other modifications are with regards to the depth of the CNN and LSTM as they increase the number of convolutional layers to four and add one more recurrent layer. They expect their proposed method to outperform established methods in sound event detection. The evaluation results confirm that expectations are met as they show an improvement in performance for the proposed method (i.e., CRNN) compared to previous approaches to sound event detection. Previous approaches, using the same dataset, report error rates between 16.0 to 17.5, while the CRNN achieves an error rate of 11.3.

A CRNN which includes a GRU network as RNN architecture is also used for music classification [6]. They compare the performances of the proposed network to three existing CNN models, and expect that the ability to capture segment structure and the flexibility introduced by the RNN benefits the classification performances. Results show an AUC score of 0.86 for the CRNN, while the AUC scores for the CNN models vary from 0.83 to 0.855. They conclude that the CRNN effectively learns deep features which could be used for prediction tasks such as predicting music tags.

Hence, the CRNN approach provides promising results across various audio classification tasks. However, according to Pishdadian et al. [31] it would be premature to disregard traditional feature representations in favour of exclusively employing deep audio features as the handcrafted features could provide the model with additional information that could not be captured by deep networks. Moreover, Kuncheva et al. [22] argues that combining complementary and diverse features could improve the classification performances of a model. Giannakopoulos et al. [12] provides support for this approach as they find a significant increase in performance for classifying urban audio events and environmental audio sounds when HaF are introduced into the models compared to models that exclusively relied on deep audio features extracted using a CNN. Giannakopoulos et al. [12] apply the different feature representations to similar Support Vector Machine (SVM) architectures to compare their performances, and reports accuracy levels of 44.2% and 52.2% for the model including exclusively deep audio features and the model combining deep audio features with HaF, respectively. They conclude that, based on a simple classification scheme

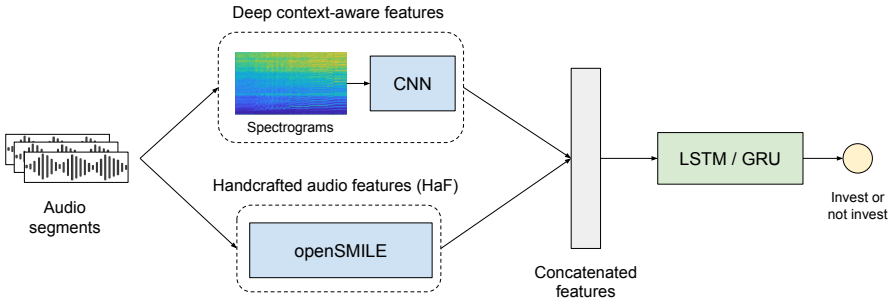

Fig. 1. Pipeline of the proposed approach. Audio segments from the pitcher are used as input to (i) pretrained VGGish network to obtain deep context-aware features and (ii) OpenSMILE to obtain Handcrafted audio Features (HaF). Deep and handcrafted features are concatenated and given as input to a LSTM or a GRU. The model predicts whether the investors will invest in the pitcher's business idea or not.

(i.e., SVM), the contextual knowledge of the input data could be significantly increased when deep audio features are combined with typical HaF. Besides, they recommend that future research should explore a similar approach in the context of a deep learning framework.

This study builds upon the recommendation for future research of Giannakopoulos et al. [12] by introducing HaF into the deep network. To apply the current state-of-the-art approach, this study uses a CRNN architecture. As GRU and LSTM networks are both commonly used in audio classification research, this study compares the performances of these two networks.

3 Method

The proposed network (see Figure 1) takes a sequence of audio segments as input, extracts deep and handcrafted audio features and concatenates them, finally passes them through a recurrent neural network (LSTM or GRU) to predict whether the startup will get investment or not.

3.1 Feature extraction

Within this study, two types of features are incorporated: (i) deep context-aware features, and (ii) Handcrafted audio Features (HaF).

Deep context-aware features The deep context-aware features are extracted with the VGGish architecture which is pre-trained on the YouTube-8M dataset [43]. For the audio input, a short-time Fourier transform with a step size of 10 milliseconds is applied on 25 milliseconds windows to compute the spectrogram. Spectrograms are mapped to 64 Mel-spaced frequency bins and frames with 96

x 64 pixels are obtained [17]. Finally, log transform is applied to obtain log Mel spectrograms, which are used as input to the pretrained VGGish model.

In this study, the VGGish network with 11 layers is used to extract deep discriminative features. Moreover, the last max-pooling layer and the last group of convolutional layers are dropped, resulting in a VGGish network architecture of four modules [43], [31]. The VGGish network outputs a one-dimensional feature vector with 128 extracted features for every ∼1-second segment of the input audio.

Handcrafted audio Features (HaF) The HaF are extracted using the open-source OpenSMILE [10], [9] toolkit. OpenSMILE extracts both Low-Level Descriptors (e.g. pitch, energy, Mel Frequency Cepstral Coefficients) as well as their functionals (e.g. extreme values, means, peaks, moments). We used LLDs and functionals in feature set eGeMAPSv02 [8]. Previous work shows the usage potential of the features from OpenSMILE in recognition and classification tasks across multimedia research [9], [36], [35]. Moreover, the ComParE feature set is applied on various multimedia tasks such as determining emotion from music, speech and sound, and delivered state-of-the-art accuracy [42]). This configuration was shown to be the best performing acoustic feature set for personality impression prediction [15].

3.2 Modeling long-term temporal information

To model the long-term information in an audio sequence, extracted features are passed through a recurrent neural network. We used two different RNNs namely, LSTM and GRU. GRU contains less parameters compared to LSTM, and generally performs well on limited training data [1] whereas LSTM can remember longer sequences [7]. Our dataset is small in terms of the number of videos and contains rather long sequences. For that reason, we used both GRU and LSTM to model long-term temporal information. We used single-layer GRU and LSTM networks and varied the number of hidden units within the set $[64, 128, 256, 512]$. The representation obtained from LSTM or GRU is passed to a binary classification layer to predict the investment decision label assigned to the corresponding audio sequence.

4 Experimental setup

4.1 Dataset

The dataset used in this study was collected for the scientific purpose of investigating how our understanding of the decision-making process, that involves social interactions in the entrepreneurial context, could be advanced by leveraging modern data science techniques [25]. Our data [24] includes video recordings from 43 individuals who perform an entrepreneurial pitch about their start-up business idea and three judges assessing the pitches. From the total of 43 pitches,

26 were performed in an offline setting while 17 were performed in an online environment. The pitchers were all students who took part in the pitch competition as part of their university-level educational program on data science and entrepreneurship. Pitchers had a maximum of three minutes to perform their pitch followed by an interactive Q&A session in which the judges, who were all experienced within the start-up ecosystem, could ask questions to the pitcher for a maximum of ten minutes. For the purpose of this study, only the audio recordings from the pitches are considered.

4.2 Preprocessing of audio recordings

Audio recordings from the pitch segments were extracted. As it is preferred to process equally sized inputs to optimize the learning process of deep learning algorithms, audio segments of 150 seconds are used as input. To this end, the audio segments for pitches shorter than 150 seconds were (partially) duplicated whereas for the longer pitches the 150 seconds from the middle of the pitch were selected.

In order to train recurrent neural networks, we segmented audio segments into non-overlapping chunks. In audio processing literature common values of segment size vary from 1 to 10 seconds [20, 12]. Earlier works often adopt a one second time frame when sounds are involved, but prefer a longer time frame (i.e., 2-10 seconds) when music or speech is involved. In this study, we set the chunk size to 2 seconds as it is reasonable to assume that speech lasts for at least 2 seconds [11].

From these 150-second audio segments we created 2-second chunks to approximate the duration of speech [11]. This approach resulted in 75 non-overlapping chunks of 2 seconds for all 43 pitches which were fed into the feature extractor. The data was split into a training (80%) and test (20%) set to evaluate the performance of the model.

4.3 Outcome predictor of investment decisions

After each pitch, each investor evaluated the performance of the pitcher and assigned a score between 0 and 100 with intermediate steps of 5 indicating the probability of investment. We mapped these scores into binary target labels for each participant consisting of *invest* (i.e., class 1) or *not invest* (i.e., class 0). Although the potential investors (i.e., judges) all had experience within the start-up business environment, their level of experience and expertise with regards to new venture start-ups, new market developments, and new product developments varied. Moreover, the potential investors had various backgrounds (e.g., venture capitalists and business coach) and field of interests (e.g., sustainability, technology, lifestyle, sports, non-profit). According to [38] and [30], the investment decision-making process, and thereby the ultimate decision, is influenced by the level of experience in the specific type of setting. In other words, two investors with experiences in different markets might evaluate a start-up business differently, resulting in a different decision on whether or not to provide

financial resources. These differences in judgment also occur within the dataset as for example one judge evaluated the probability to invest with a score of 80, while another judge assigned a score of 20 to the same pitch. Considering the diverse panel of judges, and thus the differences in evaluations, and by keeping in mind the real-world setting where new ventures are looking for at least one investor, the binary class label *invest* was assigned based on at least one positive evaluation. That is, the label *invest* is assigned when at least one of the potential investors evaluated the pitch with a probability to invest score of 50 or higher, and the label *not invest* is assigned when all scores were lower than 50.

4.4 Models

The proposed models consist of a combination of deep and handcrafted features and a RNN architecture (i.e., GRU or LSTM). In order to see the impact of handcrafted features, we performed an ablation study and used only deep features in combination with LSTM or GRU. We trained the following four models:

Model 1: LSTM with deep context-aware features and HaF (CNN + HaF + LSTM): this model includes a combination of features extracted using VGGish and OpenSMILE and an LSTM for temporal information processing.

Model 2: GRU with deep context-aware features and HaF (CNN + HaF + GRU): this model includes a combination of features extracted using VGGish and OpenSMILE and an GRU for temporal information processing.

Model 3: LSTM with deep context-aware features (CNN + LSTM): this model includes features extracted using VGGish architecture and an LSTM for temporal information processing.

Model 4: GRU with deep context-aware features (CNN + GRU): this model includes features extracted using VGGish architecture and a GRU for temporal information processing.

In general, the process of the four defined models is similar. For each pitch, 75 non-overlapping 2-seconds audio chunks are sequentially put through the feature extractor(s) which outputs a 2-dimensional feature vector for each pitch. The feature vectors of all pitches are stacked together and form a 3-dimensional input vector for the RNN, which is either the LSTM or the GRU.

We input 2-second audio chunks into VGGish network, which outputs a matrix of size 2×128 for each chunk. The post-processing process flattens the feature vector of the chunk to a 1-dimensional vector of size 1×256, while stacking all chunks together. This iterative process results in a feature matrix of size 75×256 for each pitch. A similar iterative process is defined for all files in the dataset, resulting in the 3-dimensional feature vector which is fed into the RNN (i.e., LSTM or GRU). This process is applied while developing Model 3 and Model 4.

In Model 1 and Model 2, features extracted using VGGish and OpenSMILE are concatenated to capture both deep context-aware features as well as HaF.

The OpenSMILE toolkit extracts 113 features (when LLDs and functionals are considered) for each input file resulting in a 1-dimensional feature vector of size 1×113 for each 2-second audio segment. This feature vector is concatenated with the feature vector from the VGGish network into a feature vector of size 1×369 for each chunk. While processing all the chunks, the resulting feature vectors of the chunks are stacked together in a feature matrix of size 75×369 for each file. A similar iterative process is defined for all files in the concerned dataset (i.e., training set or test set), resulting in the 3-dimensional feature vector which is fed into the RNN (i.e., LSTM or GRU).

4.5 Training

We performed hyperparameter tuning on each model based on a limited grid search. The explored hyperparameters are number of units (64, 128, 256), dropout rate (0, 0.1, 0.2), learning rate (1e-2, 1e-3, 1e-4), and number of epochs (10, 20, 50). Furthermore, each model includes the Adam optimization algorithm, which is a robust yet computationally efficient stochastic gradient-based optimization that combines the ability to deal with sparse gradient with the ability to deal with non-stationary settings [21]. Moreover, since the models are designed for a binary classification problem, the binary cross-entropy loss function was implemented in all models.

5 Results

We compare the performances of the four models (CNN + HaF + LSTM, CNN + HaF + GRU, CNN + LSTM, and CNN + GRU) to understand the impact of different RNN architectures and to analyze the impact of including HaF on the performance. Table 1 summarizes the performances of the four models.

Table 1. Performances on the test set across the proposed models. The highest performances are presented in **bold**.

	Model	Accuracy	AUC
Model 1	CNN+HaF+LSTM	**0.778**	**0.775**
Model 2	CNN+HaF+GRU	**0.778**	0.750
Model 3	CNN+LSTM	0.667	0.650
Model 4	CNN+GRU	0.667	0.625

5.1 Comparison of different RNN architectures

We compare the performances of model pairs containing the same features, but model the temporal information with different RNNs (Model 1 vs. Model 2,

and Model 3 vs. Model 4). As shown in Table 1, a similar accuracy of 66.7% is reported for the models containing only deep features regardless of the RNN architecture. Moreover, the proposed models containing a combination of deep and handcrafted features show a similar accuracy of 77.8% regardless of the RNN architecture. With regards to the AUC score, small differences between the models implementing the LSTM network sequence descriptor and models implementing the GRU network as sequence descriptor are found. The LSTM baseline model yields an AUC value of 0.650 whereas the GRU baseline model results in an AUC value of 0.625. A similar difference is observed for Models 1 and 2 as the LSTM and GRU report AUC values of 0.775 and 0.750 respectively. Hence, while in terms of accuracy scores similar performances between the LSTM and GRU network are found, differences in AUC scores are reported where models implementing the LSTM network appear to perform slightly better than models implementing the GRU network.

5.2 Impact of HaF on performance

In order to evaluate the impact of HaF, the performances of models with the same RNN, but with different features (with and without HaF) are compared. As shown in Table 1, Model 3 (CNN+LSTM) reports an accuracy score of 66.7% while Model 1 (CNN+HaF+LSTM) yields an accuracy score of 77.8%. A similar increase in performance is found for the models including the GRU network as RNN, where the models without (Model 4) and with (Model 2) HaF report accuracy scores of 66.7% and 77.8%, respectively. In terms of AUC scores, an overall 0.125 increase is reported for models including HaF. Model 3 and Model 1 result in AUC values of 0.650 and 0.775, respectively. Similarly Model 4 yields 0.625 AUC while Model 2 yields 0.750 AUC score. Hence, an increase in both accuracy and AUC scores is found when HaF are introduced into the model irrespective of the type of RNN. This could indicate that the HaF capture complementary information that benefits the model in learning and recognizing acoustic patterns.

6 Conclusion

We aim to examine to what extent an investors' decision to provide financial resources could be predicted based on vocal behavior during entrepreneurial pitches by fusing deep context-aware features and Handcrafted audio Features (HaF) in combination with a Recurrent Neural Network (RNN) architecture, particularly LSTM or GRU. Results show that models that combine deep and HaF outperform the ones without HaF, which indicates that HaF provide the models with additional information that could not be captured by deep features, and that benefits the sequential modeling performance. Moreover, this study shows that GRU and LSTM networks provide comparable performances on audio data.

This study concludes that it is promising to use vocal behavior to predict an investors' decision on whether or not to provide funding. One limitation is that we used a combined set of offline and online (recorded during the Covid-19 lockdown) pitches to have a larger amount of data to train our deep learning models. Considering the fact that non-verbal behavioral cues such as non-content characteristics of speech could be different in online and offline settings, future work could focus on investigating vocal behavior to predict investment in different settings separately. Another limitation is that we focused on predicting the binary variable reflecting whether the judges would invest in this business idea derived from the probability of investment variable. As judges do not actually make investments at the end of the competition, probability of investment may not necessarily be the most genuine assessment made by them. In future work, additional variables including originality, quality, and feasibility of the business idea could also be predicted to gain a better understanding of decision-making in an entrepreneurial context. Finally, in this work we focused only on vocal behavior. Future research on entrepreneurial decision-making based on deep learning approaches should examine the influence of combining multiple behavioral cues (e.g., facial expressions and body movements) as this could provide us with valuable insights into the way we, as humans, make decisions.

References

1. Bermant, P.C., Bronstein, M.M., Wood, R.J., Gero, S., Gruber, D.F.: Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Scientific reports **9**(1), 1–10 (2019)
2. Bonaccio, S., O'Reilly, J., O'Sullivan, S.L., Chiocchio, F.: Nonverbal behavior and communication in the workplace: A review and an agenda for research. Journal of Management **42**(5), 1044–1074 (2016)
3. de Bont, T.: Social Signal Processing in Entrepreneurial Research: a Pilot Study. Ph.D. thesis, Tilburg University (2020)
4. Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T.: Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(6), 1291–1303 (2017)
5. Carlson, N.A.: Simple acoustic-prosodic models of confidence and likability are associated with long-term funding outcomes for entrepreneurs. International Conference on Social Informatics pp. 3–16 (2017)
6. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2392–2396. IEEE (2017)
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
8. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., et al.: The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE transactions on affective computing **7**(2), 190–202 (2015)
9. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 835–838 (2013)

10. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings of 18th ACM International Conference on Multimedia pp. 1459–1462 (2010)
11. Gallardo-Antolín, A., Montero, J.M.: Histogram equalization-based features for speech, music, and song discrimination. IEEE Signal processing letters **17**(7), 659–662 (2010)
12. Giannakopoulos, T., Spyrou, E., Perantonis, S.J.: Recognition of urban sound events using deep context-aware feature extractors and handcrafted features. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 184–195. Springer (2019)
13. Goss, D.: Schumpeter's legacy? interaction and emotions in the sociology of entrepreneurship. Entrepreneurship Theory and Practice **29**(2), 205–218 (2005)
14. Grahe, J.E., Bernieri, F.J.: The importance of nonverbal cues in judging rapport. Journal of Nonverbal behavior **23**(4), 253–269 (1999)
15. Gürpinar, F., Kaya, H., Salah, A.A.: Multimodal fusion of audio, scene, and face features for first impression estimation. In: 2016 23rd International conference on pattern recognition (ICPR). pp. 43–48. IEEE (2016)
16. Han, K., Wang, D.: Neural network based pitch tracking in very noisy speech. IEEE/ACM transactions on Audio, Speech, and Language Processing **22**(12), 2158–2168 (2014)
17. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (icassp) pp. 131–135 (2017)
18. Huang, L., Pearce, J.L.: Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. Administrative Science Quarterly **60**(4), 634–670 (2015)
19. van de Kamp, L.: Predicting Investors' Investment Decisions by Facial Mimicry. Ph.D. thesis, Tilburg University (2020)
20. Kim, H.G., Moreau, N., Sikora, T.: MPEG-7 audio and beyond: Audio content indexing and retrieval. John Wiley & Sons (2006)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine learning **51**(2), 181–207 (2003)
23. Lee, H., Pham, P., Largman, Y., Ng, A.: Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in neural information processing systems **22**, 1096–1104 (2009)
24. Liebregts, W., Urbig, D., Jung, M.M.: Survey and video data regarding entrepreneurial pitches and investment decisions. Unpublished raw data (2018-2021)
25. Liebregts, W., Darnihamedani, P., Postma, E., Atzmueller, M.: The promise of social signal processing for research on decision-making in entrepreneurial contexts. Small Business Economics **55**(3), 589–605 (2020)
26. Lim, H., Park, J., Han, Y.: Rare sound event detection using 1d convolutional recurrent neural networks. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop. pp. 80–84 (2017)
27. Luo, Y., Chen, Z., Hershey, J.R., Le Roux, J., Mesgarani, N.: Deep clustering and conventional networks for music separation: Stronger together. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 61–65. IEEE (2017)

28. van Mil, C.: Improving Your Pitch with Facial Action Units... Is It Possible? Ph.D. thesis, Tilburg University (2020)
29. Milkman, K.L., Chugh, D., Bazerman, M.H.: How can decision making be improved? Perspectives on Psychological Science **4**(4), 379–383 (2009)
30. Moritz, A., Diegel, W., Block, J., Fisch, C.: Vc investors' venture screening: the role of the decision maker's education and experience. Journal of Business Economics pp. 1–37 (2021)
31. Pishdadian, F., Seetharaman, P., Kim, B., Pardo, B.: Classifying non-speech vocals: Deep vs signal processing representation. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019) (2019)
32. Pollack, J.M., Rutherford, M.W., Nagy, B.G.: Preparedness and cognitive legitimacy as antecedents of new venture funding in televised business pitches. Entrepreneurship Theory and Practice **36**(5), 915–939 (2012)
33. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 4580–4584. IEEE (2015)
34. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal processing letters **24**(3), 279–283 (2017)
35. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al.: The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France (2013)
36. Schuller, B.W.: The computational paralinguistics challenge [social sciences]. IEEE Signal Processing Magazine **29**(4), 97–101 (2012)
37. Shepherd, D.A.: Multilevel entrepreneurship research: Opportunities for studying entrepreneurial decision making (2011)
38. Slovic, P.: Psychological study of human judgment: Implications for investment decision making. The Journal of Finance **27**(4), 779–799 (1972)
39. Stoitsas, K., Onal Ertugrul, I., Liebregts, W., Jung, M.M.: Predicting evaluations of entrepreneurial pitches based on multimodal nonverbal behavioral cues and self-reported characteristics. In: Companion Publication of the 2022 International Conference on Multimodal Interaction (2022)
40. Tianyu, Z., Zhenjiang, M., Jianhu, Z.: Combining cnn with hand-crafted features for image classification. In: 2018 14th IEEE International Conference on Signal Processing (ICSP). pp. 554–557. IEEE (2018)
41. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using deep convolutional recurrent network. IEEE International Conference on Acoustics, Speech and Signal Processing pp. 5200–5204 (2016)
42. Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M., Scherer, K.R.: On the acoustics of emotion in audio: what speech, music, and sound have in common. Frontiers in psychology **4**, 292 (2013)
43. Yu, H., Chen, C., Du, X., Li, Y., Rashwan, A., Hou, L., Jin, P., Yang, F., Liu, F., Kim, J., Li, J.: TensorFlow Model Garden. https://github.com/tensorflow/models (2020)