Deciphering Entrepreneurial Pitches: A Multimodal Deep Learning Approach to Predict Probability of Investment

Pepijn van Aken
Utrecht University
Utrecht, The Netherlands
p.w.o.vanaken@students.uu.nl

Merel M. Jung
Tilburg University
Tilburg, The Netherlands
m.m.jung@tilburguniversity.edu

Werner Liebregts
Jheronimus Academy of Data Science
's-Hertogenbosch, The Netherlands
w.j.liebregts@tilburguniversity.edu

Itir Onal Ertugrul
Utrecht University
Utrecht, The Netherlands
i.onalertugrul@uu.nl

ABSTRACT

Acquiring early-stage investments for the purpose of developing a business is a fundamental aspect of the entrepreneurial process, which regularly entails pitching the business proposal to potential investors. Previous research suggests that business viability data and the perception of the entrepreneur play an important role in the investment decision-making process. This perception of the entrepreneur is shaped by verbal and non-verbal behavioral cues produced in investor-entrepreneur interactions. This study explores the impact of such cues on decisions that involve investing in a startup on the basis of a pitch. A multimodal approach is developed in which acoustic and linguistic features are extracted from recordings of entrepreneurial pitches to predict the likelihood of investment. The acoustic and linguistic modalities are represented using both hand-crafted and deep features. The capabilities of deep learning models are exploited to capture the temporal dynamics of the inputs. The findings show promising results for the prediction of the likelihood of investment using a multimodal architecture consisting of acoustic and linguistic features. Models based on deep features generally outperform hand-crafted representations. Experiments with an explainable model provide insights about the important features. The most predictive model is found to be a multimodal one that combines deep acoustic and linguistic features using an early fusion strategy and achieves an MAE of 13.91.

KEYWORDS

Multimodal interaction; Entrepreneurial pitch competition; Decision making process; Social signal processing; Nonverbal behavior

ACM Reference Format:
Pepijn van Aken, Merel M. Jung, Werner Liebregts, and Itir Onal Ertugrul. 2023. Deciphering Entrepreneurial Pitches: A Multimodal Deep Learning Approach to Predict Probability of Investment. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3577190.3614146

ICMI '23, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614146

1 INTRODUCTION

Delivering a successful elevator pitch on a business proposal in front of investors is an intimidating challenge for many entrepreneurs. Convincing investors of the potential of a business plan and raising funds to realize the plan are critical parts of the entrepreneurial process. We are only beginning to understand how investors make decisions regarding investments [5].

Decision-making, the process of determining the most appropriate course of action based on available information, in the field of entrepreneurship is characterized by high levels of uncertainty [25]. Berner et al. [2] claim that the main premise of entrepreneurship is to accept high levels of risk while investing or producing a good. Decision makers in entrepreneurial contexts usually have to rely on heuristics as factual information is often lacking or limited [13]. The high levels of uncertainty and the use of heuristics make it difficult to analyze the decision-making process, and due to its complexity, this process has attracted a lot of academic research (e.g. [25, 29]). It is especially interesting to study decisions that involve social interactions, since due to the lack of factual information, this social interaction itself can influence the decision.

The interaction between a pitching entrepreneur and an evaluating investor is such an entrepreneurial setting based on social relationships that is marked by high uncertainty. The investor has to make an assessment of the feasibility of a project based on the content of the pitch and financial data. However, research suggests that investors also rely on subtle social cues that they extract from the pitch. Huang and Pierce [13] have found that investors both rely on intuition and formal analysis when making this decision. Furthermore, they indicate that this intuition is for a large share based on the perception of the founding entrepreneur. Multiple studies have shown that this perception of the entrepreneur is shaped by verbal and non-verbal cues in the pitch. For example, the use of language and storytelling was found to play a key role in entrepreneur-investor interactions [20], non-verbal behavior cues have been found to influence the perceived passion of an entrepreneur [4], and the use of a combination of verbal cues and hand gestures was shown to have a strong positive effect on funding decisions [5].

Processing the verbal and non-verbal cues emitted by pitchers could open up a valuable source of information for research into investors' decision-making. To unlock the potential of these signals, natural language processing can be used to analyze the cues in the use of language in the pitch, while the social signal processing domain provides the tools to automatically code nonverbal signals, resulting in a more accurate and efficient analysis [16]. However, as noted by [5], the effect of verbal and non-verbal communication strategies are often studied in isolation (e.g., [10]). Since social interactions are the interplay of verbal and non-verbal cues, integrating them in a single analysis could provide interesting new insights. Vocal behavior has been identified as a potential important driver for investments in entrepreneurial pitches [5, 13], but has not been studied in a combined model with verbal cues yet. This is a current gap in existing literature which will be explored in this study.

Extracting informative features from the raw data is one of the main challenges when using acoustic and linguistic data. Traditionally, both of these modalities have been studied using hand-crafted feature sets. For the acoustic features of speech, features such as pitch and loudness can directly be extracted from the audio signal and used to make predictions such as emotion classification (e.g., [18, 19]. For language, a Bag-of-Words (BoW) model can be used to create a vector that represents the input text. Despite the fact that hand-crafted feature sets have been successfully applied on a number of tasks, there are some limitations regarding this approach, such as modeling the context of a text. The development of deep learning enabled the creation of models that can learn to extract feature representations themselves. Furthermore, in combination with deep feature embeddings, deep encoders can capture the temporal dynamics of the signals, leading to better performance [30]. Currently, these deep learning based feature extractors have become state-of-the-art in audio and language research. However, this does not imply that hand-crafted feature sets are not useful anymore (e.g., see [10]). Elbanna et al. [7] and Johnson and Marcellino [14] argue that using an ensemble of hand-crafted and deep feature sets can lead to an increase in performance and interpretability of a model. To properly analyze the role of vocal and verbal cues in decision-making in entrepreneurial contexts, both have to be considered in a single model, resulting in a multimodal approach. In a multimodal model, different unimodal models are combined and thus it captures a wider range of behavior. Different techniques exist to fuse unimodal models into a multimodal model, such as early and late fusion [22].

In this work, we propose a multimodal approach that combines both acoustic and linguistic features to predict the likelihood of investment of entrepreneurial pitches. Hand-crafted acoustic features are exacted using openSMILE [9] and deep acoustic features using the VGGish convolutional neural network [12]. As Gated Recurrent Unit (GRU) contains less parameters compared to Long Short-Term Memory network (LSTM), and generally performs well on limited training data, we fed these representations into a GRU that learns temporal dynamics from the audio signal. In addition, hand-crafted linguistic features are extracted using Linguistic Inquiry and Word Count (LIWC) [3] and deep linguistic features using Longformer [1]. Performance of early and late fusion approaches are compared by either first combining the different feature sets or combining the models' predictions, respectively. Our results show that a multimodal approach that combines the best performing features for each modality using an early fusion strategy yields the best performance. We also train an explainable multimodal model to provide insights into feature importance, that was found to perform slightly worse that the ones trained with deep representations. By performing cross-domain experiments, we show that our multimodal model developed for in-person pitches can generalize well to a different context consisting of online recordings. Our findings could provide insights for investors and researchers into the decision-making process and help entrepreneurs enhance their pitching performance.

Overall, this study contributes to existing literature in three ways:

- We propose the first multimodal approach that aims to model verbal and nonverbal behavior during social interactions in an entrepreneurial context to predict the probability of investment.
- We developed an explainable multimodal model and show that despite a slight reduction in the performance, it enables us to investigate the features that play an important role in determining the probability of investment
- We perform cross-domain experiments to show that our models that are trained using in-person recordings generalize reasonably well to online settings.

2 METHODS

2.1 Entrepreneurial pitch data-set

In this study we used the *Entrepreneurial Pitch dataset* containing video recordings from entrepreneurial pitch competitions [17]. The dataset includes video recordings of entrepreneurial pitches by university students with accompanying Q&A sessions, as well as survey data from investor judges. The data set consists of 42 pitches recorded from 2018 to 2021. The survey data contains the individual assessments of the members of the investor panel, including the probability that they would invest in the pitched business idea on a scale from 0-100. The data collection and management process has been approved by the university's ethics board. Informed consent was obtained from the pitchers and investors for their data to be used for research.

Due to the COVID-19 pandemic starting within this period, pitch sessions were conducted both in-person ($n = 25$) and online ($n = 17$). We use in-person recordings to perform within-domain experiments. Kuhn and Sarfati [15] found that the move to online settings may have affected investors' perception of social signals, with acoustic features playing a more substantial role in online settings. To investigate how our models generalize to online settings, we perform cross-domain experiments and use online recordings as test set.

2.2 Data pre-processing

Several data pre-processing steps were completed before feature extraction and training of the models: (i) extracting audio from video data, (ii) trimming the audio data, (iii) splitting the audio data in chunks, (iv) converting speech to text and (v) creating a single score for likelihood to invest.

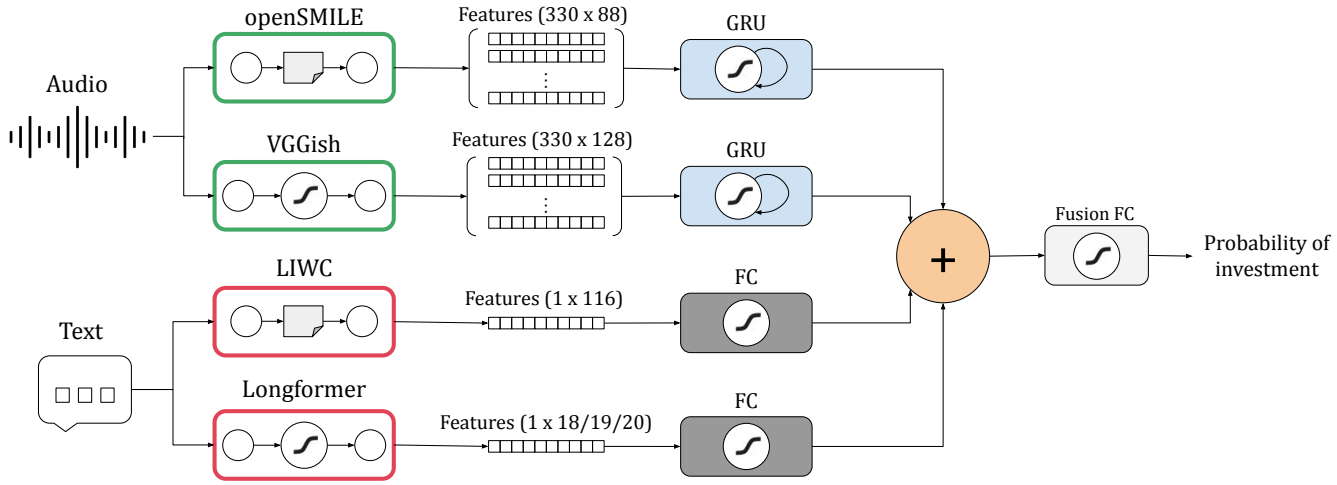

Figure 1: Multimodal method to predict the probability of investment. The inputs from different modalities are passed through specific data representation modules, in this case, openSMILE and VGGish for the acoustic, and LIWC and Longformer for the linguistic modality. Then the output representations of these modules are passed through either a GRU (acoustic features) or fully-connected (linguistic features) layer. Finally, the representations are selected, fused, and the probability of investment is predicted.

2.2.1 Extracting audio from video data. In this study, only acoustic and linguistic features are considered, thus the videos are first converted into audio files. The use of WAV format has been chosen due to its uncompressed nature, which allows for the preservation of more information for feature extraction using the openSMILE and VGGish packages.

2.2.2 Trimming the audio data. In order to accurately analyze the pitches, it is necessary to trim the audio files to only include the pitch segment itself and exclude the Q&A session. We have manually extracted the pitch portion of the audio from the full video, as the pitch may not always begin at the start of the video or end at the three-minute mark.

2.2.3 Splitting the audio data in chunks. VGGish provides a feature embedding for every 0.96 seconds of audio data converted to a log-mel-scale spectrogram. To have a fair comparison, we aim to have hand-crafted openSMILE features to represent same duration. We split all the audio data in non-overlapping chunks of 0.96 seconds to avoid repetitive information.

2.2.4 Converting speech to text data. Before linguistic features can be extracted, the audio files have to be transcribed to text data. This is done by using Google's Speech-to-Text API, which has state-of-the-art accuracy on automatic speech recognition tasks.

2.2.5 Creating a single score for likelihood to invest. Every pitch in the data set is evaluated by several judges, resulting in multiple scores per pitch. To train the model, we choose the highest score given by any judge as the output label. Our motivation is that a pitch that excites one investor, even if not others, is likely to be successful in raising money. Additionally, investors may lack expertise in the presented industry and their low score may not reflect the pitch's true quality.

2.3 Feature Extraction

2.3.1 Acoustic features. For the acoustic modality, we extracted two categories of feature sets: explainable, hand-crafted features extracted using openSMILE and deep features extracted using VGGish.

OpenSMILE: Out of the openSMILE kit [9] we specifically use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPSv02) feature set [8]. This is a basic standard acoustic parameter set intended to provide a common baseline for research. It consists of 88 features, including Low Level Descriptors (LLDs) and functionals, which are extracted from the 0.96 second chunks of audio created in the pre-processing. These features are organized into a $T \times 88$ dimensional matrix for each pitch, where T represents the number of 0.96 second chunks that fit into the length of the pitch. As audio features will be fed into a GRU which requires all inputs to be the same size, we ensure that all pitches have fixed-length feature vectors. We fix the length of all pitches based on the size of the second longest pitch ($T = 330$). This allows for the incorporation of as much audio information as possible while accounting for the longest pitch (09:25 minutes), which may be considered an outlier. To create equal length feature vectors, shorter pitches are zero padded and the longer pitch is trimmed. This is a standard method to create equal length features (e.g., [11]) in the audio modality. In addition, a "static" openSMILE representation is obtained for the entire audio recording for training the explainable model, that is not fed into GRU for learning the temporal dynamics of the audio signal. We extract the *eGeMAPSv02* (Eyben et al., [8]) features over the whole pitch at once. Consequently, a one-dimensional vector of length 88 is obtained for the entire pitch.

VGGish: VGGish converts every 0.96 seconds of audio input into a semantically meaningful 128-D embedding [12]. When feeding the pitches into the VGGish framework this results in a $T \times 128$

embedding for every pitch, where T is the number of chunks. Here, the same procedure has been applied to fix the length, by zero padding and trimming all videos to a size 330×128, resulting in deep representations of the audio chunks that are fed into the GRU.

2.3.2 Linguistic features. For the linguistic modality, hand-crafted features are obtained using LIWC while the deep features are obtained using Longformer.

LIWC: LIWC is a text analysis tool that determines the percentage of words in a text that fall into one or more linguistic, psychological and topical categories. The core of the tool is a dictionary containing words that belong to these categories. The most recent version, LIWC-22, is used to extract 116 features for each pitch [3]. These features are also used to train an explainable model, which enables us to draw conclusions on what word "categories" play a role in the investment decision-making process.

Longformer: Most transformer based models cannot be used to provide representations for long text sequences. When the textual representation of an entire pitch is desired, models such as BERT are not feasible as they provide representations for much shorter text (e.g., up to 512 words). For this reason, we used Longformer, which has a linear (instead of a quadratic) attention mechanism, allowing much longer input texts [1]. Using the output of the pooled layer of the Longformer model, a 768-dimensional embedding is obtained for the text of every pitch. Given the large number of features in the Longformer model compared to the LIWC model, we apply principal component analysis (PCA) to reduce the number of features. For every fold in the training process, principal components are obtained only using the training set, then the coefficients are used to also transform the test set.

2.4 Proposed multimodal approach

We propose a multimodal approach that fuses acoustic and linguistic information to predict probability of investment (see Figure 1). In this pipeline acoustic features are extracted using openSMILE and VGGish for a sequence of chunks. Then, they are fed into GRU to model the temporal dynamics of acoustic behavior. Linguistic features that are extracted using LIWC and Longformer for the whole text are fed into a fully connected layer with ReLU activation. By performing unimodal experiments we identify the best acoustic (openSMILE or VGGish) and linguistic (LIWC or Longformer) features. Finally, we concatenate the best performing acoustic and linguistic features and feed them into a final fully-connected layer to predict investment.

3 EXPERIMENTS

3.1 Ablation studies

In order to evaluate the performance of individual components and alternative design choices, we perform an ablation study. In the first set of ablation experiments, we investigate the performance of using individual feature sets represented in four individual branches in Figure 1. We perform experiments with individual types of features (openSMILE, VGGish, LIWC, and Longformer). Acoustic features are fed into GRU and then to the output layer. Linguistic features are first fed into a fully connected layer and then to the output layer.

In the second set of ablation experiments, we investigate the impact of combining hand-crafted and deep features within a modality. We perform experiments with early and late fusion strategies. For the early fusion of acoustic modality, we concatenate the representations obtained from openSMILE and VGGish and then feed them into a single GRU. For the late fusion, we fuse the decisions of the regressors trained only with openSMILE and only with VGGish features by taking their average. For the linguistic modality we concatenate representations obtained from LIWC and Longformer for the early fusion whereas in late fusion we take the average of the predictions of regressors trained with LIWC and Longformer separately. Note that these experiments provide unimodal performances considering either acoustic or linguistic features.

In the third set of ablation experiments, we combine multimodal features in two ways: (1) combine the best performing features for each modality, and (2) combine all features (all of the four branches). We perform experiments with early and late fusion strategies. While computing late fusion results, decisions of regressors trained with individual sets of features are fused by taking their average for all test instances during inference.

3.2 Explainable multimodal model

To achieve an explainable model, we combine static openSMILE and LIWC features. The eGeMAPSv02 feature set is used to extract an 88-dimensional acoustic feature vector for the entire pitch. 116-dimensional linguistic feature vectors are obtained for each pitch using LIWC. The final multimodal vector is derived by concatenating these feature vectors, resulting in a 204-dimensional vector. We use Xgboost Regressor as our model. We have performed grid search over the number of instances, maximum depth, and learning rate. We use the SHAP (SHapley Additive exPlanations) framework to estimate Shapley values [24], which were introduced in game theory to gauge each player's participation in cooperative games. These values are used to gain insights into which features played an important role in predicting the likelihood of investment, and to analyze the impact of acoustic and linguistic features on investment likelihood.

3.3 Cross-domain experiments

Models that perform well when trained and tested within the same domain may not generalize well to unseen domains. To evaluate the generalizability of our models that are trained on in-person videos, we evaluate the best performing ones on the 17 pitches recorded in the online setting. We refer to the online setting as cross-domain as nonverbal communication is limited and eye-to-eye contact is missing. We extract acoustic and linguistic features from the online videos following the same procedure as in the initial experiments. We evaluate performance of the best performing acoustic, linguistic, and multimodal models on all instances of the online dataset.

3.4 Training and evaluation

The 25 in-person recorded pitches included in this study were recorded over four distinct sessions, each with different pitchers and investors. The pitches are divided into four folds, with each fold consisting of all the pitches from a single session, in order to ensure that each fold or test set can be considered representative

of an actual session. Hyper-parameter tuning is conducted using a grid search with 5-fold cross-validation for number of instances, learning rate, and maximum depth when Xgboost Regressor is used. Deep learning models are trained using a grid search to optimize the hyperparameters namely number of units in the GRU layer, learning rate, and dropout. We used Adam optimizer and early stopping is applied using a validation split of 0.2. Batch size is set at 5 for all models. Performance of the models is evaluated using the Mean Absolute Error (MAE) and feature importance is based on SHAP scores.

4 RESULTS

Table 1 shows the performances of the proposed multimodal approach and its individual unimodal components. The table includes the Mean Absolute Error (MAE) for each individual fold, as well as the average MAE across all folds.

4.1 Unimodal models

The results show that among acoustic features, VGGish outperforms openSMILE yielding an average MAE of 15.41. For the linguistic features, LIWC and Longformer perform similarly but Longformer slightly outperforms LIWC with an average MAE of 15.60. The best performing acoustic model outperforms the linguistic models, with the VGGish model yielding the best results overall. For both modalities the models that use deep representations outperform the hand-crafted interpretable feature sets.

Combining deep and hand-crafted acoustic features with an early fusion strategy (14.82 MAE) yields better performance compared to their late fusion (MAE = 16.16). On the contrary, combining deep and hand-crafted linguistic features with a late fusion strategy (MAE = 14.85) outperforms early fusion (MAE = 18.44). When we compare these results with individual feature sets, we can see that combining deep and hand-crafted features outperform models using only deep or only hand-crafted features for both acoustic and linguistic modalities. Overall, the best performing unimodal model is obtained when VGGish and openSMILE features are combined with an early fusion strategy, yielding an MAE of 14.82.

4.2 Multimodal models

We present the results of experiments with multimodal models and different fusion strategies. The best performing set of features of each modality are combined, namely VGGish for the acoustic modality and Longformer for the linguistic modality using either early or late fusion strategies. The lowest MAE (13.91) is obtained using early fusion. Model performance was also evaluated for the combination of all four feature sets. Results show that combining all features decreased performance (MAE = 15.22) compared to only combining the best performing feature sets of each modality.

4.3 Explainable multimodal model

The results show that the explainable multimodal model, that combines the hand-crafted acoustic (openSMILE) and linguistic (LIWC) features performs worse (MAE = 15.59) than the multimodal models that combine either the best performing feature sets or all features. However, we note that the improvement of this model in terms of explainability, compared to the models containing deep representations, is obtained at a relatively small cost in terms of model performance. The difference between the performance of the best performing model (early fusion of the VGGish and Longformer representations) and the explainable model is rather limited (average MAE of 13.91 versus 15.59). At the cost of this slight decrease in performance, we do gain a lot of interesting insights by looking at the SHAP feature importance plots of these models.

It is most interesting to examine the distributions of the feature categories (acoustic or linguistic) over the feature importance plots obtained for the 4 different folds (see Figure 2). Across all the 4 different models both acoustic and linguistic are found amongst the most predictive features. Looking at the overall distribution, we find this is skewed to the linguistic features, since 60 out of the 80 most important features across the four models are of the linguistic category. However, when only considering the top 5 features of the 4 models, the number of acoustic and linguistic features is very similar (11 versus 9). Furthermore, in 3 models, the strongest feature is an acoustic one. This finding demonstrates that when developing a multimodal model consisting out of acoustic and linguistic features, important features used by the model to make predictions originate from both modalities.

We also identify the explainable features that appear in the top 20 of feature importance for at least three out of the four folds. The 6 common features consist of 4 linguistic (*Clout*, *Conversation*, *Word Count*, and *Number*) and 2 acoustic features (*F0semitoneFrom27.5Hz pctlrange0-2* and *F2frequency amean*).

The linguistic feature *Clout* is defined as the "language of leadership" [3]. A higher number for the Clout score indicates that the presenter is speaking from a perspective of high expertise and is confident, on the other hand, lower scores suggest a more tentative or humble speaking style [21]. For this feature, the SHAP plots for three of the four models indicate that a higher value for this feature would lower the output of the model. This would mean that when the presenter seems to be highly confident, this has a negative impact on getting an investment in the pitch setting studied here.

The acoustic feature *F0semitoneFrom27.5Hz pctlrange0-2* is a measure for the distribution of the fundamental frequency over the audio signal and is useful to identify extremes [6]. In folds 2, 3, and 4, a lower value of this feature has a positive impact on the investment probability.

We also find the acoustic feature *F2frequency amean* in the list of common important features. The second formant (F2) frequency is related to the vowel sounds of speech. It shows up as the most predictive feature in one of the models and is found important in three out of four folds. Higher values of this feature have a positive impact on the probability of investment score.

The linguistic feature *Word Count (WC)* represents the number of words used during the pitch. It is found in the top-20 features of all four models. However, based on the findings it is difficult to infer the impact of word count on the probability of investment.

For the linguistic feature *Conversation*, higher values of this feature are observed to have a substantial negative impact on the investment score. Even though the pitch setting only involves a single speaker rather than a bidirectional interaction, this category involves non-fluent speech such as: 'oh', 'um' and 'uh', and also

Model	Features	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	Average MAE
Acoustic	openSMILE	18.54	15.23	13.44	20.92	17.03
Acoustic	VGGish	15.48	12.42	11.53	22.19	15.41
Linguistic	LIWC	17.20	16.17	11.13	18.01	15.63
Linguistic	Longformer	16.03	12.25	10.94	23.21	15.60
Acoustic (EF)	VGGish + openSMILE	14.22	17.53	10.38	17.13	14.82
Acoustic (LF)	VGGish + openSMILE	16.80	13.98	12.48	21.39	16.16
Linguistic (EF)	LIWC + Longformer	14.49	19.56	15.02	24.70	18.44
Linguistic (LF)	LIWC + Longformer	16.62	12.61	9.56	20.61	14.85
Multimodal (EF)	VGGish + Longformer	17.17	13.51	5.47	19.47	**13.91**
Multimodal (LF)	VGGish + Longformer	15.11	11.73	10.47	22.47	14.95
Multimodal (EF)	All	16.48	15.21	9.17	20.52	15.35
Multimodal (LF)	All	16.49	12.94	10.76	20.70	15.22
Multimodal (EF)	Explainable: openSMILE + LIWC	13.92	13.54	12.56	22.36	15.59

Table 1: Regression results evaluated using mean absolute error (MAE) to predict probability of investment using different feature sets. Feature sets are combined using either early fusion (EF) or late fusion (LF) strategies. MAE 1 to 4 indicate the results for the model evaluated on hold-out pitch sessions 1 to 4 respectively.

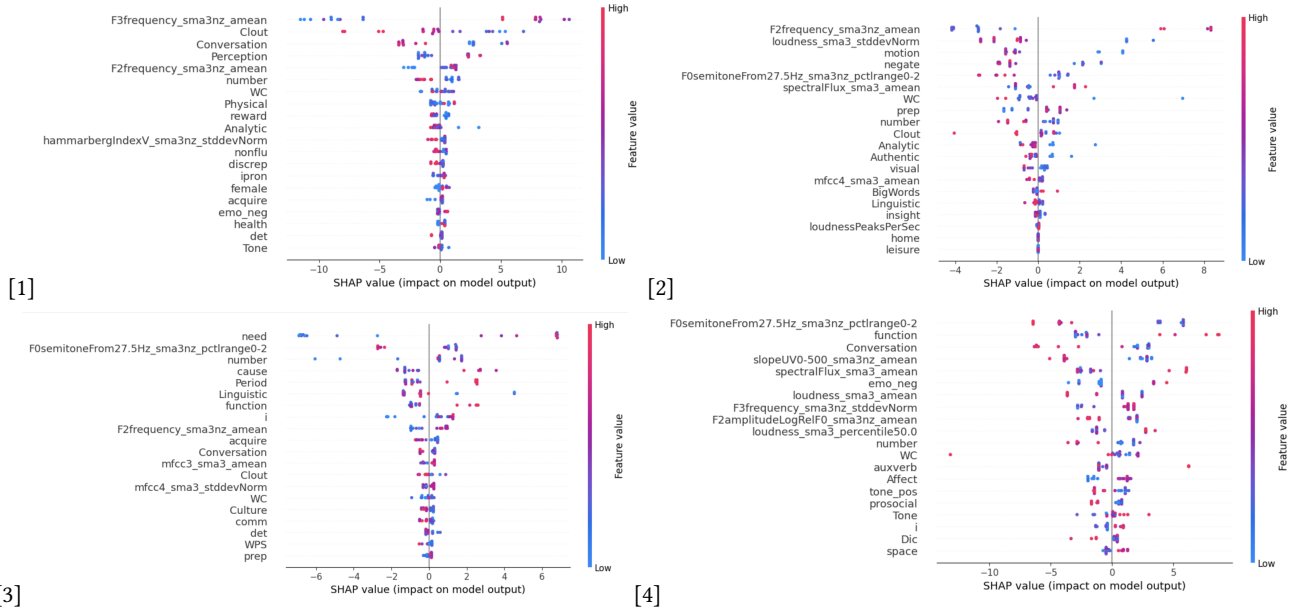

Figure 2: SHAP summary plots for the multimodal models, one for each training fold. The summary plot combines feature importance with feature effects. Features are ordered according to their importance on the y-axis and the position on the x-axis is determined by the Shapley value. The color represents the value of the feature from low to high.

contains filler words. Therefore, the SHAP values for the *Conversation* feature indicate that when a pitcher presents with a lack of fluency, for example caused by stammering or usage of filler words, this has a negative impact on the likelihood of investment.

The linguistic feature *Number* represents a count of the use of numbers in the text. SHAP values in Figure 2 suggest that a lower value of the feature *Number* has a positive impact on the probability of investment. This indicates that in the pitch setting studied here, using a lot of numbers during the pitch could have a decreasing effect on the probability of investment.

4.4 Cross-domain experiment

Comparing the result of the cross-domain experiments to those of the within-domain experiments suggests that the models generalize

Modality	Model	*MAE 1*	*MAE 2*	*MAE 3*	*MAE 4*	Average MAE
Acoustic	VGGish + openSMILE (EF)	16.51	15.91	19.05	21.21	18.17
Linguistic	LIWC + Longformer (LF)	19.15	19.50	21.65	19.62	19.98
Multimodal	VGGish + Longformer (EF)	18.59	16.22	16.91	16.84	**17.14**

Table 2: Cross-domain results of models trained on the in-person pitches and evaluated on the online pitches

to an online pitch settings to a certain extent (see Table 2). While a slight decrease in performance across all three models is observed, the size of this decrease is relatively small and the models continue to exhibit an adequate performance. Notably, some of the patterns observed in the previous experiments recur in the cross-domain experiment. Firstly, when comparing the unimodal acoustic and linguistic models, the acoustic model is again found to be superior. The difference in performance is more pronounced in this cross-domain experiment than in the within-domain experiment, where these particular acoustic and linguistic models show relatively similar scores. Therefore, we observe that the features representing the acoustic modality generalize better to the online setting than the linguistic features. Secondly, like in the within-domain experiments, the multimodal model was found to outperform the individual unimodal models. The relative increase in performance compared to the acoustic model consisting of concatenated features fed into a single GRU is comparable to the in-person context results, namely a decrease in MAE score of around 1.0.

5 DISCUSSION AND LIMITATIONS

5.1 Discussion of the results

The aim of this study was to examine the decision-making processes of investors during entrepreneurial pitch interactions by analyzing the acoustic and linguistic characteristics of these pitches. Previous work has shown that early-stage investors rely on two main components when making decisions: factual data on the viability of the project and perceptions of the founding entrepreneur [13]. Given the scarcity of factual data in this area, the resulting decisions are often characterized by a high degree of uncertainty. Therefore, the perception of the entrepreneur, for a large part based on the social interaction between the investor and entrepreneur, forms a vital part of the decision-making process.

As an ablation study, we studied a set of unimodal models, some consisting of a single feature set and others consisting of a combination of feature sets. Across all the unimodal acoustic and linguistic models, we note that the best performing acoustic model is the one that combines VGGish and openSMILE feature representations and feeds them into a GRU. Similarly, the best performing linguistic model is identified as the one that combines LIWC and Longformer models by averaging the output of the predictions of individual models. These findings suggest that, in both cases, a combined set of hand-crafted and deep features forms the optimal representation of a modality. The acoustic model slightly outperforms the linguistic model (MAE of 14.82 compared to 14.85), but this difference is negligible.

The experiments on multimodal models provide three notable insights. Firstly, comparing a multimodal model using all features to a model using only the best performing features showed that the latter approach performs better in both early and late fusion architectures. In a multimodal context the combination of both hand-crafted and deep features outperformed models based on only one of the feature sets. This finding is not consistently replicated in the multimodal scenario but is consistent with the previous work [26, 28] on multimodal architectures.

The second insight concerns the comparison between early fusion and late fusion models. When using the best features only, the early fusion model outperforms the late fusion model. However, when all feature sets are used, the late fusion model is superior. There is no clear consensus on the optimal fusion strategy, as previous work has yielded mixed results [22]. Therefore, it is recommended to experiment with both early and late fusion strategies to assess their impact on performance. The early fusion approach learns a comprehensive feature space, while the late fusion approach is straightforward to implement and applicable in a broader range of contexts.

Thirdly, multimodal models have been compared to unimodal models to analyze if they outperform them. One non-explainable multimodal model, which used early fusion of VGGish and Longformer features, was found to be superior to the best performing unimodal models. This suggests that a multimodal approach is a viable methodology for studying investment decision-making based on pitches, and highlights the added value of studying different modalities in conjunction.

Explainable models have been developed in order to examine what features play an important role when the models predict the likelihood of investment. For these models, openSMILE and LIWC features are used. First we look at the model performance of these explainable models. Compared to the best performing non-explainable models, we observe a slight decrease in performance. The acoustic explainable model (MAE of 16.82) performs worse than the acoustic model were a GRU is used to model the openSMILE and VGGish features at once. Similarly, the explainable multimodal model is outperformed by all the four other multimodal models. However, we observe that this decrease in performance is rather limited and at the cost of this decrease we do gain a lot in terms of explainability. Furthermore, we again observe that the multimodal model outperforms the unimodal models indicating that also in the context of explainable models, a multimodal architecture is most suitable to predict the likelihood of investment.

Shap feature importance values were obtained to identify the most discriminating features and to study correlations between feature values and model predictions. Overall, a relatively large overlap was found between the most important features across multiple models. LIWC measures four broader "summary" variables and two of these, *Clout* and *Analytic*, were found to be common important features across the four linguistic models. In the analysis

of the openSMILE features, 3 different functionals over the fundamental frequency were common important features. In addition, both the mean and the normalized standard deviation of the loudness were also frequent relevant features. For the SHAP plots of the multimodal models, the primary focus was on examining the distribution of modality categories across the feature importance plots. The results of the analysis demonstrate that both acoustic and linguistic features play an important role in determining the model output. This finding further supports the argument that the combination of verbal and non-verbal behavioral cues captures a more comprehensive range of behavior and subsequently yields stronger predictors.

We can also compare our results to previous work reporting the performance of regression models developed based on the in-person recordings available in this data set [23, 27]. These authors also used the probability of investment score as a target variable and used different feature sets of nonverbal behavioral cues from several modalities such as facial expressions, head movement and vocal expressions. The best performing model by [23] was trained on a combination of action unit features and deep facial expression representations and yielded an average MAE of 17.80 whereas [27] achieved an average MAE of 16.47 with their model trained on head movement features. Our strongest model achieves a state-of-the-art performance of 13.91, yielding an improvement over previous work.

5.2 Limitations

Although the presented findings show promising results for the prediction of investor's likelihood of investment, some limitations and context need to be considered.

A possible confounding factor is that the analyzed pitches were given by university students as part of their university program. Even though the pitches were judged by a panel of professional investors, the investors would not invest actual money. Also, the evaluating investors had limited information on the economic viability of the pitches, which is an important factor in investment decisions. As a result, this could have increased the impact of the pitch delivery on the investment probability. Thus, our models may be less generalizable to more realistic entrepreneurial contexts. Moreover, due to the relatively small dataset of 25 in-person pitches (and 17 cross-domain test pitches), the presented models were based on 18 to 20 training instances, possibly affecting the models' generalizability. Fortunately, the dataset will be expanded as yearly pitch sessions are organized in the context of university courses on data science startups, allowing to train more robust models. Additionally, pitch data will be collected from students enrolled in a graduate program specifically aimed at entrepreneurship, hereby increasingly mimicking a professional entrepreneurial context.

Moreover, it should be noted that our models have been trained on the audio recordings of the pitches whereas the question and answering (Q&A) sessions with investors afterwards might have also influenced the investors' decision. Future work could analyze the Q&A session to develop a more comprehensive model. Additionally, in real-world entrepreneurial settings, social interactions between entrepreneurs and investors can be more extensive and involve multiple meetings, hereby broadening the scope of the interaction beyond what was captured in this study.

Finally, some considerations should be noted regarding the explainable models and the effect of common important features on the outcome of the models. SHAP can enhance predictive models by revealing associations between features and outcomes. However, interpreting the values as specific features that can be manipulated to change predictions is often misleading as correlation does not imply causation. While the SHAP tool provides a method to create transparency regarding correlations, it does not indicate causation. Therefore, if the goal is to create guidelines for entrepreneurs to enhance their pitching skills, it is essential to exercise caution when interpreting feature importance plots in this context and it would be necessary to conduct additional causal analyses.

6 CONCLUSION AND FUTURE WORK

In this work, we propose a multimodal approach that combines acoustic and linguistic features extracted from recordings of entrepreneurial pitches to predict the likelihood of investment. Both modalities are represented using hand-crafted and deep features. GRU, a deep learning model has been used to model the temporal dynamics of the inputs. The acoustic and linguistic models have been combined in a single multimodal pipeline by applying early and late fusion of the feature representations. Moreover, explainable models trained on the hand-crafted features were developed to identify and interpret important features.

Our findings show promising results for the prediction of investor's likelihood of investment of entrepreneurial pitches using acoustic and linguistic features. State-of-the-art performance has been achieved on this dataset using a multimodal model where the best performing features of each modality are integrated using an early fusion strategy. In the experiments, deep features generally outperform hand-crafted ones. Further findings suggest that when developing a unimodal model, it is beneficial to represent this modality using both hand-crafted and deep feature sets. It was found that early fusion outperforms late fusion. Across multiple explainable models, consistent features were found to be important predictors. A cross-domain experiment demonstrated that the developed models for in-person pitches generalize to an online setting to some extent.

In conclusion, this study has demonstrated that using a multimodal analysis approach is a promising direction for studying decision-making in the context of entrepreneurial pitches. Based on this, future research in this area could continue to build upon the proposed methodology to address some of its limitations. An alternative direction for future research could involve expanding the methodology used in this study. For instance, it would be of interest to incorporate the visual modality, which includes features such as facial expressions, gestures, and head movement into the multimodal analysis.

REFERENCES

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).
[2] Erhard Berner, Georgina Gomez, and Peter Knorringa. 2012. 'Helping a large number of people become a little less poor': The logic of survival entrepreneurs. *The European Journal of Development Research* 24, 3 (2012), 382–396.
[3] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* (2022).

[4] Xiao-Ping Chen, Xin Yao, and Suresh Kotha. 2009. Entrepreneur passion and preparedness in business plan presentations: a persuasion analysis of venture capitalists' funding decisions. *Academy of Management journal* 52, 1 (2009), 199–214.

[5] Jean S Clarke, Joep P Cornelissen, and Mark P Healey. 2019. Actions speak louder than words: How figurative language and gesturing in entrepreneurial pitches influences investment judgments. *Academy of Management Journal* 62, 2 (2019), 335–360.

[6] William E Cooper and John M Sorensen. 2012. *Fundamental frequency in sentence production.* Springer Science & Business Media.

[7] Gasser Elbanna, Alice Biryukov, Neil Scheidwasser-Clow, Lara Orlandic, Pablo Mainar, Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. 2022. Hybrid Handcrafted and Learnable Audio Representation for Analysis of Speech Under Cognitive and Physical Load. *arXiv preprint arXiv:2203.16637* (2022).

[8] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia.* 1459–1462.

[10] Ilona Goossens, Merel M Jung, Werner Liebregts, and Itir Önal Ertuğrul. 2022. To invest or not to invest: Using vocal behavior to predict decisions of investors in an entrepreneurial context. In *International Conference on Pattern Recognition.* Springer, 273–286.

[11] Wenjing Han, Tao Jiang, Yan Li, Björn Schuller, and Huabin Ruan. 2020. Ordinal learning for emotion recognition in customer service calls. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 6494–6498.

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp).* IEEE, 131–135.

[13] Laura Huang and Jone L Pearce. 2015. Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. *Administrative Science Quarterly* 60, 4 (2015), 634–670.

[14] Christian Johnson and William Marcellino. 2022. Bag-of-Words Algorithms Can Supplement Transformer Sequence Classification & Improve Model Interpretability. (2022).

[15] Nicole Kuhn and Gilberto Sarfati. 2021. Zoomvesting: angel investors' perception of subjective cues in online pitching. *Journal of Entrepreneurship in Emerging Economies* (2021).

[16] Werner Liebregts, Pourya Darnihamedani, Eric Postma, and Martin Atzmueller. 2020. The promise of social signal processing for research on decision-making in entrepreneurial contexts. *Small business economics* 55, 3 (2020), 589–605.

[17] Werner .J. Liebregts, Diemo. Urbig, and Merel M Jung. 2018-2021. Survey and video data regarding entrepreneurial pitches and investment decisions. *[Unpublished raw data]* (2018-2021).

[18] Iker Luengo, Eva Navas, Inmaculada Hernáez, and Jon Sánchez. 2005. Automatic emotion recognition using prosodic parameters. In *Ninth European conference on speech communication and technology.* Citeseer.

[19] Erik Marchi, Florian Eyben, Gerhard Hagerer, and Björn W Schuller. 2016. Real-Time Tracking of Speakers' Emotions, States, and Traits on Mobile Platforms.. In *INTERSPEECH.* 1182–1183.

[20] Martin L Martens, Jennifer E Jennings, and P Devereaux Jennings. 2007. Do the stories they tell get them the money they need? The role of entrepreneurial narratives in resource acquisition. *Academy of management journal* 50, 5 (2007), 1107–1132.

[21] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report.

[22] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[23] Arya Tri Prabawa, Merel M Jung, Kostas Stoitsas, Werner Liebregts, and Itir Önal Ertuğrul. 2022. Predicting Probability of Investment Based on Investor's Facial Expression in a Startup Funding Pitch. In *Proceedings of BNAIC/BeNeLearn 2022.*

[24] Lloyd S Shapley. 1997. A value for n-person games. *Classics in game theory* 69 (1997).

[25] Dean A Shepherd, Trenton A Williams, and Holger Patzelt. 2015. Thinking about entrepreneurial decision making: Review and research agenda. *Journal of management* 41, 1 (2015), 11–46.

[26] Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal analysis and estimation of intimate self-disclosure. In *2019 International Conference on Multimodal Interaction.* 59–68.

[27] Kostas Stoitsas, Itir Önal Ertuğrul, Werner Liebregts, and Merel M Jung. 2022. 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.* 121–126.

[28] Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction.* 406–413.

[29] Nigel Wadeson. 2006. Cognitive aspects of entrepreneurship: decision-making and attitudes to risk. *The Oxford handbook of entrepreneurship* (2006).

[30] Gang Zhou. 2021. Donation-Based Crowdfunding Title Classification Based on BERT+ CNN. In *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing.* 291–296.