# Understanding Users' Confidence in Spoken Queries for Conversational Search Systems

Youjing Yu[1], Zhengxiang Shi[2], and Aldo Lipani[2]

[1] University of Cambridge, Cambridge, United Kingdom
yy471@cam.ac.uk
[2] University College London, London, United Kingdom
{zhengxiang.shi.19,aldo.lipani}@ucl.ac.uk

**Abstract.** The confidence level in users' speech has long been recognised as an important signal in traditional dialogue systems. In this work, we highlight the importance of user confidence detection in *queries* in conversational search systems (CSSs). Accurately estimating a user's confidence level in CSSs is important because it enables the CSSs to infer the degree of competency of a user on the queried topic and subsequently tailor its responses appropriately. This is especially important in CSSs since their responses need to be concise and precise. However, few prior works have evaluated user confidence in CSSs due to a lack of available datasets. We present a novel speech-based dataset named UNderstanding Spoken qUeRiEs (UNSURE) [3], which contains confidence grading annotations of user queries in natural language conversations. Based on this dataset, we propose a multimodal approach to infer users' confidence in spoken queries as a baseline model. Preliminary experimental results demonstrate that our proposed fusion model is capable of achieving near human-level performance.

**Keywords:** user understanding · conversational search · conversational system

## 1 Introduction

Conversational search systems (CSSs) enable users to engage in mixed-initiative interactions by expressing their queries in natural language and interacting with the system through multiple turns of exchanges [15,24,9]. Compared to the use of written language, the use of speech in CSSs can make these interactions more convenient, as the use of hands is not required. However, the amount of information that CSSs can convey at any given turn is limited. This makes it crucial to present the user with not just the most relevant answer, but also an answer that is appropriate for the user's level of understanding and expertise [28,25].

Brennan and Williams [1] suggested that successful communication requires the person to accurately estimate and monitor not only this person's own knowledge state but also the knowledge state of their conversational partners. Inferring

[3] Code and instructions on how to obtain this dataset is available at https://github.com/YoujingYu99/confidence_css

how well someone else knows something is termed as the Feeling of Another's Knowledge, or FOAK [1]. The listener's perception of a speaker's expressed confidence or FOAK allows the listener to infer the knowledge state of the speaker. If the speaker raises a query, the listener's perception of the speaker's confidence level in this query is important for the listener to infer about the speaker's knowledge state and subsequently make a decision on how and what to reply [19]. This is backed by many studies which find that there exists a positive relationship between confidence in a person's voice and their expertise in the field of the conversational topic [6,18,21]

In the context of CSSs, where the listener is the system, a successful evaluation of the perceived confidence level in the user's voice enables CSSs to obtain a deeper understanding of the user's desires, beliefs, and intentions as well as the situational and conversational context, and thus the search can be enhanced. For example, in the context of CSSs for education, if a user asks the system: "What does hypothesis mean?" and the system deems the user's confidence level to be high, the system may hence infer that the user has some expertise in the field of statistics and proceed to suggest the formal definition: "a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation." In contrast, if the system deems the user's confidence level to be low, the system may infer that the user has limited prior knowledge in the field of statistics and suggest a simpler explanation: "a possible explanation that may not be correct."

Despite its importance, user confidence in CSSs has not been systematically studied, with few works assessing user confidence in conversations. The major challenge is the lack of appropriate datasets. A summary of the available datasets that focus on human confidence detection in speech is presented in Table 1. The existing datasets are mostly collected from Question Answering (QA) conversations. They focus on confidence detection in the speech of the person *answering* the question, which is mostly of assertive or declarative nature. On the other hand, there is no available dataset that specifically focuses on queries or interrogative speech. To tackle the aforementioned issues, we propose a new dataset named UNSURE, which stands for UNderstanding Spoken qUeRiEs. This dataset is based on the Spotify Podcast Dataset [3], which contains queries raised by either the interviewer or the interviewee during podcast sessions and their corresponding confidence scores annotated by Amazon Mechanical Turk (MTurk) workers. In addition, We propose a fusion (speech and text) model to predict the confidence scores as a baseline model for this task.

The rest of the paper is organised as follows: Section 2 outlines the related works in the field of emotion recognition in speech and confidence assessment. Section 3 describes the methodologies for the collection, labelling and processing of the dataset. Section 4 describes the confidence prediction task, and Section 5 describes the fusion model. Section 6 presents the results and discussions of the proposed fusion model. Finally, we conclude in Section 7.

**Table 1.** Summary of the statistics of the existing datasets and our proposed dataset **UNSURE**.

| | Size | Raters | Score Range | Annotations Performed on | Other Limitations |
|---|---|---|---|---|---|
| Nair et al. [19] | 254 | 3 | 0 - 100 | Self-introductions | Size too small |
| Chanda et al. [2] | 1 242 | 2 | 3 Categories | Audiovisual answers | Number of raters |
| Pon-Barry [23] | 1 700 | 6 | 5 Categories | Answers to handwritten digits | Limited Topics |
| Martin et al. [16] | 956 | - | 0-4 | Answers to colour names | Limited Topics |
| **UNSURE** | 4 542 | 3 | 0 - 5 | Queries | - |

## 2 Related Work

Nair et al. [19] conducted an initial investigation into the prediction of speakers' confidence and evaluated the performance of different neural network structures on this task. Their dataset consists of two parts: the first is based on recordings of university students' self-introductions, and the second is based on an audio dataset from Kaggle. Chanda et al. [2] also created an audiovisual dataset based on a collection of interviews, which contains the interviewees' answers to questions raised by interviewers. Pon-Barry [23] prepared The Harvard Uncertainty Speech Corpus, where participants view a train route illustration and answer a question about the train timings. The images of the digits presented to participants are taken from the handwritten MNIST dataset and vary in terms of ambiguity and legibility, representing a measure of certainty about the identity of the digit. However, the Harvard Uncertainty Speech Corpus is limited in scope to the topic of digits and does not provide a comprehensive understanding of confidence in the context of CSSs.

Another existing dataset is prepared by Martin et al. [16], where participants were asked to name the colour they saw, and the entropy level associated with each colour was calculated based on the number of unique names that participants have given each colour. Apart from the limited range of conversational topics, we point out another very important limitation here: while the calculation of confidence based on the number of names each colour has been called is a clear and well-defined approach for the domain of colour, it is very difficult to generalise this approach to other domains where the context is much more complicated. Most importantly, all of the aforementioned datasets are recordings of conversations, which mainly contain declarative speech rather than questions or queries. Hence their use is limited in CSSs, as the emphasis is placed on queries raised by users.

## 3 The UNSURE Dataset

We propose a new dataset built upon the Spotify Podcast Dataset, named UNSURE. The UNSURE dataset contains a total of 4 542 audio samples, each of which is one query sentence. We now describe the steps we took to generate this dataset.

### 3.1 Dataset Collection

***Pre-processing***. We first pre-process the Spotify Podcast Dataset with its word-level transcription files. Firstly, we selected podcasts within certain categories: Business, Government, History, News, Science, Religion & Spirituality, Society & Culture and Technology. This is done to make sure to cover a wide range of topics and to select queries that the users are more likely to ask CSSs. From the transcription files, we then *filter out* the podcasts and sentences based on the following criteria. At a podcast level, we removed (1) podcasts with only one speaker since we prefer conversations and not monologues; (2) podcasts that are clean of interjecting sounds (e.g., "hmm", "oh", etc.) since the absence of these sounds is a sign of edited podcasts; we want unedited podcasts to capture the authenticity of speech. This is backed by the study by Brennan and Williams [1] that confirmed that a listener's perceived confidence in the speaker is largely affected by the presence of fillers and interjecting sounds. At a sentence level, we removed (1) sentences which contain more than one speaker tag since multiple speakers in one audio clip may affect the consistency of the tonal quality when used for training and (2) sentences which are shorter than 2 seconds as these are deemed too short to provide enough context.

***Annotation Process***. A total of 7 919 audio clips are sampled and uploaded to MTurk for crowd-workers to grade the confidence of the audio clip from 0 to 5, with an interval of 0.5 (0 = very unconfident, 5 = very confident). Each MTurk Human Intelligence Task (HIT) is annotated by three crowd-workers. In the instructions section of the HIT, three sample audio clips are provided for reference, which the authors of this paper deem to have a confidence score of 0, 2.5 and 5. In the answers section, workers rate the confidence level of the speaker. In addition, they have to answer two questions: the first (Q1) to identify whether the speaker in the audio clip is asking a question and the second (Q2) to identify whether there are multiple speakers in the audio. These two questions are also used to verify the answers of the crowd workers.

Out of the 7 919 audio clips, 219 samples of audio clips are from the verification set we prepared for quality control purposes. The authors of this paper listened to each of the sample clips and gave answers to Q1 and Q2. Each HIT consists of 12 audio clips to which the crowd-worker gives a score and answers Q1 and Q2 for each clip. Two out of the 12 clips are from the verification set. Hence, we set a threshold of 50%, and workers who answered more than 50% of the verification questions wrong are deemed to be giving random answers instead of truthful answers and hence rejected. The rejected HITs are then republished for new workers to work on. After seven rounds of rejection and republishing, we collect the results, which are now deemed truthful.

Having gathered data for 7 919 audio clips, using Q1 and Q2, we filter out the clips which are not questions, contain more than one speaker and are assessed by fewer than 3 workers. The final dataset consists of 4 542 audio clips. Since the workers are given example audio clips and the associated confidence scores, and only results from workers deemed truthful are accepted, no normalisation is done

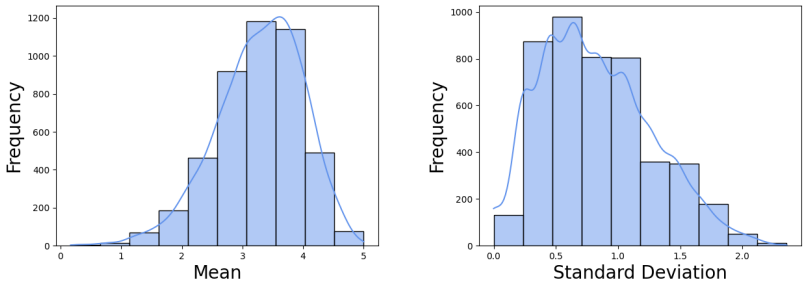


**Fig. 1.** Histograms for mean scores (left) and standard deviation (right) of each audio.

on the scores the workers gave to avoid introducing potentially artificial agreements. Table 1 compares our dataset with other datasets available for assessing user confidence. It is important to note that our dataset is not only relatively large with a suitable number of raters, but it is also entirely query-based, which is useful in the context of CSSs.

### 3.2 Dataset Analysis

***Scores Distribution.*** For the average score of three scores given for 4 542 audio samples, the exact number of scores in each category is 16, 168, 1099, 2375 and 884. Half of the samples have a score between 3 and 4 (relatively confident), while very few samples have a score between 0 and 1 (extremely unconfident). The histogram for the mean and standard deviation of the scores per audio are plotted in Figure 1. The mean for the audio scores centred around 3.5, which is slightly higher than the mean of the scoring range of 2.5. This is expected since the audio samples were extracted from podcasts; we would expect the speakers to have more confidence in their voice since the topic of conversation is usually in their domain of expertise. However, this also means that we have more samples which are fairly confident compared to samples which are not confident at all. The standard deviation for the audio follows averaged around 0.7, which means that the scores are fairly close together. However, the presence of many audio samples with a standard deviation greater than 1.0 indicates disagreement in the crowd-workers' judgement.

***Rater Agreement.*** To understand and quantify how well the crowd-workers agree with each other, we measure the inter-rater reliability [12,27] using the Intraclass Correlation Coefficient (ICC) [7], the Pearson correlation coefficient (Pearson's r) [20] and Kendall's tau [14]. Table 2 shows the agreement scores for the dataset among the original three workers. All three values are less than 0.5, indicating limited agreement among the three workers. By inspecting the results closely, we notice many outliers in the scores, that is, cases when a worker's

**Table 2.** Inter-rater reliability experiments, where ICC stands for the intraclass correlation coefficient. ICC has a range between −1 and 1 while Kendall's tau and Pearson's r have a range between 0 to 1, and a higher score indicates a better agreement.

| | **ICC** | **Pearson's r** | **Kendall's tau** |
|---|---|---|---|
| Among 3-raters | 0.181 | 0.306 | 0.040 |
| Our vs 3-raters average | 0.428 | 0.277 | 0.179 |
| Our vs closer 2-raters average | 0.339 | 0.205 | 0.117 |

score is in strong contradiction with the scores given by two other workers. We recognise two possible explanations for such a limited agreement:

1. The limited agreement among the workers is due to outliers in data, which results from the workers not giving truthful scores when rating the confidence of the speaker.
2. The limited agreement among the workers is not due to outliers in data. All scores given by the workers are truthful, and they are the result of individual confidence perception.

If the second explanation is deemed more likely, the three scores given by all three workers should be used. On the other hand, if the first explanation is deemed more likely, we should discard the outliers, which is the score furthest from the other two. Hence we only keep the two most similar scores and use the average of these two scores as the label.

We proceed to devise an experiment to test which reason is more likely. We first label a test set consisting of 492 speech samples ourselves. Note that this is the same set we used as the test set for model evaluation in Section 6. These labels are hence regarded as reliable and consistent and will be referred to as the expert validation set onwards. We then proceed to calculate the user agreement between our scores and the average between three raters (using all three scores) and two raters (discarding outliers), and the results are shown in Table 2. Interestingly, we see that the agreement between our score and the three-rater average is better than that between our score and the closer two-rater average. This suggests that the average of the three raters should be regarded as more reliable than the average between two raters. This indicates that the outliers are not a result of the workers giving random answers, and hence, we should keep all three scores and use their averages as the labels.

Through this experiment, we gain the important insight that it is inherently difficult for humans to agree much with each other on the confidence level of speech. The disagreement among raters on the confidence level of speech is likely not to result from poor quality of the data or untruthful answers provided by workers. Rather, it illustrates the difference in individual understanding of confidence assessment and the subjective nature of the hearer-centric affect labelling paradigm [23]. This is indeed very common in the labelling of subjective qualities. For instance, in the study by Flexer and Grill [8] where the participants were asked to grade whether the candidate song that was played to them

was similar to a query song, their results only achieved a Pearson correlation of 0.40. Moreover, when Erkelens et al. [5] performed research where primary care experts were asked to listen to recorded emergency calls and rate the safety of triage of patients with chest discomfort, the ICC among the medical experts was reported to be only 0.16. Furthermore, Perski et al. [22] carried out another study where raters were asked to rate the top ten features from a pre-specified list that are most important for the reduction of alcohol consumption and they found that the overall ICC is only 0.15. These strengthen our argument that the limited rater agreement is not a problem but, rather, demonstrates the subjective nature of the rating of human confidence. Our measure of averaging the scores between three workers has served to mitigate this inherent disagreement between human ratings as much as possible.

## 4 Speech Confidence Prediction Task

### 4.1 Task Definition

The objective of this task is to learn a function, denoted as $f$, that can predict the level of uncertainty of a speaker when posed with an audio question, represented as $x_s$, along with its corresponding transcription in natural language form, represented as $x_t$. The model function can be expressed as:

$$y = f(x_s, x_t) \tag{1}$$

This equation indicates that the uncertainty score, denoted as $y$, is a function of the audio $x_s$ and the transcription $x_t$. The value of $y$ is expected to fall within the range of 0 to 5.

### 4.2 Evaluation Metrics

To evaluate the model's performance on this task, we use two metrics: (1) Mean Squared Error (MSE) and (2) a parameterized accuracy rate.

On the one hand, MSE is a measure of the average squared difference between the predicted values of a model and the ground-truth values. MSE provides a quantitative measure of the performance of the model, where a lower value of MSE indicates better predictive performance.

On the other hand, the accuracy considers a prediction to be correct if the uncertainty score falls within a range of $\pm 0.5$ (inclusive) of the ground-truth value. The decision to use a tolerance value of $\pm 0.5$ is based on our observations during the re-annotation process of the expert validation set. We found that when the same sample was annotated and re-annotated after a week, the scores assigned were typically within a margin of 0.5. As a result, we determine that a margin of $\pm 0.5$ would be an acceptable level of accuracy for the predicted scores.

Formally, given the ground-truth value, denoted as $y_i$, and the predicted result, denoted as $\hat{y}_i$, the equation for the accuracy metric is computed as follows:

$$Accuracy_{0.5} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{[y_i - 0.5, y_i + 0.5]}(\hat{y}_i), \tag{2}$$

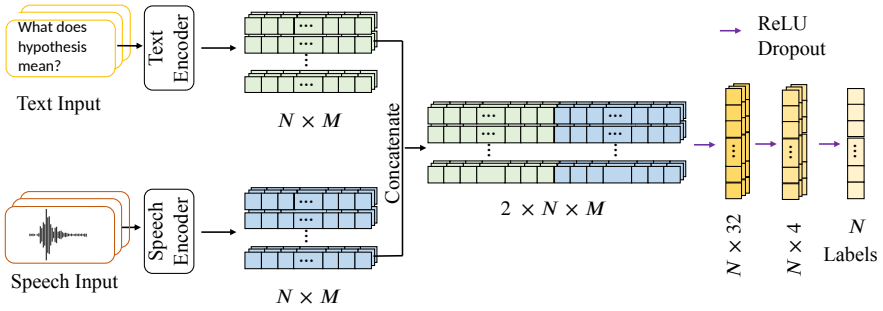


**Fig. 2.** Model architecture with three components, *text encoder*, *speech encoder* and *fusion module*. $N$ represents the batch size and $M$ stands for the dimension of the speech and text encoder outputs.

where $N$ is the sample size and $\mathbb{1}$ is the indicator function that returns a value of 1 if the predicted score $\hat{y}$ falls within a range of $\pm 0.5$ of the ground-truth score $y$, and a value of 0 otherwise.

## 5 The Proposed Method

This section outlines our proposed baseline model for the task of predicting the level of confidence exhibited by a user in response to an audio question and its corresponding transcriptions.

Our proposed model is composed of three major components: *text encoder*, *speech encoder* and *fusion module*. For the text encoder, we used the $\text{BERT}_{\text{base}}$ model [4] to convert the natural language transcription of the audio question into a contextual representation, and we specifically selected the CLS vector as the transcription representation, which typically contains 768 dimensions. Other text encoders can be used in place of BERT, but we chose BERT due to its simplicity and popularity. The resulting CLS vector is further processed through two dense layers and an activation layer as part of the text encoding process.

For the speech encoder, we utilised the pre-trained $\text{HuBERT}_{\text{base}}$ model, as outlined by Hsu et al. [13], to transform the raw audio signal of the user's speech into a continuous representation. The pooled output is further processed through two dense layers and an activation layer as part of the text encoding process.

Finally, the fusion module combines the encoded text and speech inputs to generate the predicted confidence score. The network structure for the fusion model is depicted in Figure 2, with $N$ representing the batch size. A fusion layer utilises the concatenated output from the speech encoder (dimension $M$) and the text encoder (dimension $M$). After the fusion layer, the concatenated output is processed through two dense layers and an activation layer.

**Table 3.** Test results for a random classifier, a naive average baseline classifier, experts, and our proposed fusion model. The top-performing results are highlighted in bold.

| | **Accuracy**$_{0.5}$ | **MSE** |
|---|---|---|
| Random Classifier | 0.197 | 3.160 |
| Experts | **0.416** | 1.060 |
| Fusion Model | 0.411 | **0.923** |

In general, this proposed baseline model aims to serve as a foundational framework for subsequent development in the area of predicting user speech confidence.

## 6 Experiments and Results

In this section, we first compare the proposed model performance with those of random classifiers and expert validation. Additionally, we perform an ablation study on speech-only and text-only models. Our experimental results are presented in Table 3 and 4.

### 6.1 Main Results

To evaluate the performance of our fusion model, we compare it against two baseline models: (1) a random number generator that generates continuous values randomly between 0 and 5 (inclusive); and (2) expert labels provided in the expert validation set. Experiments are conducted using the same two metrics described in Section 6.1.

***Accuracy***. The accuracy of a random number generator simulated is 0.197, which is expected given that we have a tolerance of 1.0 on a scale from 0 to 5. However, this represents the worst-case scenario, and we anticipate that the model would have an accuracy higher than 0.200 if it is able to effectively extract useful features expressing confidence level from the speech and text. Note that the expert performance achieves an accuracy of only 0.416. This is promising as our model, with an accuracy of 0.411, is close to achieving human-level performance, reaching the upper bound in confidence score prediction. We argue that this limited accuracy level achieved by the fusion model does not imply a poor performance in this scenario since even human beings rarely agree with each other on the confidence level of a speaker, as demonstrated by the low accuracy of the expert performance. Rather, this result indicates that the model is able to extract similar relevant features for confidence detection as human beings, showcasing its effectiveness and usefulness.

**Table 4.** Test results for fusion model, speech-only model and text-only model. The top-performing results are highlighted in bold font.

| | $\textbf{Accuracy}_{0.5}$ | **MSE** |
|---|---|---|
| Speech-Only Model | 0.405 | 1.015 |
| Text-Only Model | 0.274 | 1.932 |
| Fusion Model | **0.411** | **0.923** |

***MSE.*** We also calculate the MSE between the scores and ground-truth labels. The random number generator achieves the highest value for MSE, which is 3.160. The MSE between the expert performance and the labels is 1.060, and that between the model performance and the labels is 0.923. Although the model performs slightly worse in terms of accuracy compared to the expert, it performs slightly better in terms of MSE. This could be attributed to the fact that the model was directly optimised using the MSE loss function, and thus, it has learned to minimise the loss between the score and the predictions.

***Summary.*** Our experimental results have highlighted the importance of utilising both the accuracy and MSE metrics when evaluating the performance of models in predicting human confidence levels in speech. The accuracy metric provides an indication of the proportion of correct predictions made by the model, while the MSE measures the average deviation between the predicted scores and the ground-truth labels. Therefore, it is advisable to consider both metrics as performance indicators when evaluating human confidence in speech.

### 6.2 Ablation Studies

To determine which input, speech or text, contains more information on the confidence level of the individual raising the query, we conduct further experiments by training the model on speech input only and then text input only while keeping the other input encoder frozen. Our hypothesis is that speech carries more information on human confidence compared to text. This is because speech encompasses various features that better depict our confidence level, such as the vocal volume [26], vocal pace [11], duration and frequency of disfluencies such as pauses or interjecting sounds [10].

The experimental results are presented in Table 4. The results show that the fusion model, which receives both speech and text inputs, outperforms the speech-only and text-only models in terms of accuracy and error, demonstrating the importance of incorporating both sources of information. However, the speech-only model outperforms the text-only model in terms of accuracy and MSE, which supports our prior belief that speech conveys more information on human confidence than text. It is worth noting that passing speech input alone into the model can achieve good accuracy, strengthening our prior argument that speech embeds much more information on human confidence expression than text.

## 7 Conclusion and Future Work

In this work, we propose UNSURE, a novel dataset for the task of human confidence evaluation in query-based speech, to facilitate further research on speech confidence in this field of the development of information retrieval and conversational systems. Through rater agreement analysis and expert validation on the UNSURE dataset, we gain the important insight that there is a fundamental difference in the human understanding of confidence, which remains a significant challenge in the evaluation of human confidence in speech. We also presented a fusion transformer-based model as a baseline model, which achieves an accuracy of 0.411 and a low mean squared error (MSE) of 0.923, comparable to the performance of human experts.

For future work, incorporating visual elements into the confidence detection model could be considered, such as collecting a query-based video dataset that incorporates facial expressions and body language of the user, which might provide additional information on the confidence level of the user [17]. With the emergence of multimodal avatars such as Botanic Human Machines which are equipped with webcams, a combination of input modes such as text, speech and video will potentially be better able to predict user confidence levels during interactive sessions. The proposed dataset and models may also help in the development of personalised chatbots. Specifically, accurately monitoring the confidence level in the user's speech enables the chatbot to monitor the knowledge state of the user throughout the conversational sessions. Future work can hence explore the possibility of user-aware chatbots which not only dynamically tailor their responses based on the user's confidence level, but also guide the user behaviour in the turn of conversations in an appropriate manner, hence maximising expected outcome and achieving a more balanced human-chatbot relationship.

## A Appendix

***Dataset Details.*** The dataset, which contains $4,542$ instances, is split into the train, validation and test sets of size $3\,600$, 450, and 492 respectively. Given the imbalanced distribution of scores within the dataset, we employed upsampling on underrepresented instances prior to the training phase to guarantee an equal number of audio samples across all ranges. Additionally, we augmented the audio samples in the audio encoder and fusion module through random modifications to their pitch, pace, and loudness, as well as the introduction of white noise.

***Training Details.*** To find the set of hyperparameters that give the optimal performance, we validated the learning rate, the number of layers to freeze in the pre-trained $\text{BERT}_{\text{base}}$ and $\text{HuBERT}_{\text{base}}$ models and the number of dense layers. The learning rate was explored between the range $5e^{-8}$ and $5e^{-6}$. We also experimented with freezing the first 10, 11 and all layers in the pre-trained models and two numbers of dense layers were tested (2 and 3) after the fusion

**Table 5.** Results for hyperparameter searching, where the bold texts highlights the best-performing results.

| Learning Rate | Trainable Layers in BERT and HuBERT | Number of Dense Layers | Val Accuracy (%) | Val Loss |
|---|---|---|---|---|
| $5e^{-8}$ | None | 2 | 28.3 | 0.148 |
| $5e^{-8}$ | Last Hidden Layer | 2 | 40.9 | 0.080 |
| $1e^{-7}$ | None | 2 | 40.9 | 0.075 |
| $1e^{-7}$ | Last Hidden Layer | 2 | **41.1** | **0.071** |
| $1e^{-7}$ | Last Hidden Layer | 3 | 33.6 | 0.091 |
| $5e^{-7}$ | None | 2 | 40.1 | 0.079 |
| $5e^{-7}$ | Last Hidden Layer | 2 | 36.3 | 0.137 |
| $5e^{-7}$ | Last Hidden Layer | 3 | 28.2 | 0.148 |
| $5e^{-6}$ | Last Hidden Layer | 2 | 40.5 | 0.083 |
| $5e^{-6}$ | None | 3 | 39.6 | 0.096 |

layer. A total of 20 combinations of the hyperparameter settings were tested and the validation accuracy and loss for the 10 combinations which give the best results are shown in Table 5. The final model hyperparameters chosen are a learning rate of $1e^{-7}$ with a linear weight decay rate of $1e^{-9}$. We also freeze the first 11 layers of $\text{BERT}_{\text{base}}$ and $\text{HuBERT}_{\text{base}}$ models, and the two dense layers after the concatenation stage have 32 and 4 neurons respectively.

Tahn activation is used before the final layer. A dropout level of 50% is introduced after all the hidden layers and layer normalisation is applied to prevent overfitting. The number of total trainable parameters for the fusion module is hence 46 810 561. The loss function chosen is the Mean Square Error (MSE) loss. The batch size is kept at 16 and the number of epochs on average is 60. Each epoch runs for approximately 10 minutes on a 24 GB Nvidia Titan RTX.

## References

1. Brennan, S.E., Williams, M.: The feeling of anothers knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. Journal of memory and language **34**(3), 383–398 (1995), https://www.sciencedirect.com/science/article/pii/S0749596X85710170?via%3Dihub
2. Chanda, S., Fitwe, K., Deshpande, G., Schuller, B.W., Patel, S.: A deep audiovisual approach for human confidence classification. Frontiers in Computer Science **3** (2021). https://doi.org/10.3389/fcomp.2021.674533, https://www.frontiersin.org/articles/10.3389/fcomp.2021.674533/full
3. Clifton, A., Pappu, A., Reddy, S., Yu, Y., Karlgren, J., Carterette, B., Jones, R.: The spotify podcast dataset. arXiv preprint arXiv:2004.04270 (2020), https://www.researchgate.net/publication/340541821_The_Spotify_Podcasts_Dataset
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Min-

nesota (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
5. Erkelens, D.C., Rutten, F.H., Wouters, L.T., de Groot, E., Damoiseaux, R.A., Hoes, A.W., Zwart, D.L.: Limited reliability of experts' assessment of telephone triage in primary care patients with chest discomfort. Journal of Clinical Epidemiology **127**, 117–124 (2020). https://doi.org/https://doi.org/10.1016/j.jclinepi.2020.07.016, https://www.sciencedirect.com/science/article/pii/S0895435620301839
6. Favazzo, L., Willford, J.D., Watson, R.M.: Correlating student knowledge and confidence using a graded knowledge survey to assess student learning in a general microbiology classroom. Journal of microbiology & biology education **15**(2), 251–258 (2014), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4278496/#:~:text=The%20moderate%20positive%20correlation%20observed,that%20confidence%20rises%20with%20knowledge.
7. Fisher, R.A.: Statistical methods for research workers. In: Breakthroughs in statistics, pp. 66–70. Springer (1992), https://psychclassics.yorku.ca/Fisher/Methods/chap6.htm
8. Flexer, A., Grill, T.: The problem of limited inter-rater agreement in modelling music similarity. Journal of New Music Research **45**, 239 – 251 (2016)
9. Fu, X., Yilmaz, E., Lipani, A.: Evaluating the cranfield paradigm for conversational search systems. In: Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval. p. 275–280. ICTIR 22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3539813.3545126, https://doi.org/10.1145/3539813.3545126
10. Goberman, A.M., Hughes, S., Haydock, T.: Acoustic characteristics of public speaking: Anxiety and practice effects. Speech communication **53**(6), 867–876 (2011)
11. Guyer, J.J., Fabrigar, L.R., Vaughan-Johnston, T.I.: Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. Personality and Social Psychology Bulletin **45**(3), 389–405 (2019)
12. Hallgren, K.A.: Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology **8**(1), 23 (2012), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/
13. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021), https://arxiv.org/abs/2106.07447
14. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938), https://www.jstor.org/stable/pdf/2332226.pdf
15. Liu, Z., Zhou, K., Wilson, M.L.: Meta-evaluation of conversational search evaluation metrics. ACM Transactions on Information Systems (TOIS) **39**, 1 – 42 (2021), https://www.semanticscholar.org/paper/Meta-evaluation-of-Conversational-Search-Evaluation-Liu-Zhou/05e4c6e0edd230accd1976f91a6350dfd470a1ab
16. Martin, L., Stone, M., Metze, F., Mostow, J.: A methodology for using crowdsourced data to measure uncertainty in natural speech. In: 2014 IEEE Spoken Language Technology Workshop (SLT). pp. 95–99 (2014). https://doi.org/10.1109/SLT.2014.7078556, https://ieeexplore.ieee.org/document/7078556

17. Maslow, C., Yoselson, K., London, H.: Persuasiveness of confidence expressed via language and body language. British Journal of Social and Clinical Psychology **10**(3), 234–240 (1971)
18. Mudavanhu, Y., Zezekwa, N.: Relationship between confidence and knowledge of the nature of science: student-teachers perspective in zimbabwe. Young (2017), https://www.researchgate.net/publication/279853899_RELATIONSHIP_BETWEEN_CONFIDENCE_AND_KNOWLEDGE_OF_THE_NATURE_OF_SCIENCE_STUDENT-TEACHERS_PERSPECTIVE_IN_ZIMBABWE
19. Nair, S., Mohan, M., Rajesh, J., Chandran, P.: On finding the best learning model for assessing confidence in speech. 2020 The 3rd International Conference on Machine Learning and Machine Intelligence (2020). https://doi.org/10.1145/3426826.3426838, https://dl.acm.org/doi/10.1145/3426826.3426838
20. Pearson, K.: Note on Regression and Inheritance in the Case of Two Parents. Proceedings of the Royal Society of London Series I **58**, 240–242 (1895), https://royalsocietypublishing.org/doi/10.1098/rspl.1895.0041
21. Pell, M.D.: Cerebral mechanisms for understanding emotional prosody in speech. Brain and language **96**(2), 221–234 (2006)
22. Perski, O., Baretta, D., Blandford, A., West, R., Michie, S.: Engagement features judged by excessive drinkers as most important to include in smartphone applications for alcohol reduction: A mixed-methods study. DIGITAL HEALTH **4**, 2055207618785841 (2018). https://doi.org/10.1177/2055207618785841, https://doi.org/10.1177/2055207618785841, pMID: 31463077
23. Pon-Barry, H., Shieber, S.M., Longenbaugh, N.S.: Eliciting and annotating uncertainty in spoken language. In: Proceedings of the 2014 Language Resources and Evaluation Conference (2014), https://dash.harvard.edu/handle/1/12149963
24. Radlinski, F., Craswell, N.: A theoretical framework for conversational search. Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (2017), https://www.semanticscholar.org/paper/A-Theoretical-Framework-for-Conversational-Search-Radlinski-Craswell/ba3659ef1d5835c07ba0de91f61fe8c3611b3bf1
25. Salle, A., Malmasi, S., Rokhlenko, O., Agichtein, E.: Studying the effectiveness of conversational search refinement through user simulation. In: Hiemstra, D., Moens, M.F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) Advances in Information Retrieval. pp. 587–602. Springer International Publishing, Cham (2021), https://link.springer.com/chapter/10.1007/978-3-030-72113-8_39
26. Scherer, K.R., London, H., Wolf, J.J.: The voice of confidence: Paralinguistic cues and audience evaluation. Journal of Research in Personality **7**(1), 31–44 (1973)
27. Tinsley, H.E., Weiss, D.J.: Interrater reliability and agreement. In: Handbook of applied multivariate statistics and mathematical modeling, pp. 95–124. Elsevier (2000), https://www.sciencedirect.com/science/article/pii/B9780126913606500057
28. Trippas, J.R., Spina, D., Cavedon, L., Joho, H., Sanderson, M.: Informing the design of spoken conversational search: Perspective paper. In: Proceedings of the 2018 conference on human information interaction and retrieval. pp. 32–41 (2018), https://dl.acm.org/doi/abs/10.1145/3176349.3176387