Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09709-5

Introducing the 3MT_French dataset to investigate the timing of public speaking judgements

Beatrice Biancardi[1] · **Mathieu Chollet**[2] · **Chloé Clavel**[3,4]

Accepted: 17 November 2023

Abstract
In most public speaking datasets, judgements are given after watching the entire performance, or on thin slices randomly selected from the presentations, without focusing on the temporal location of these slices. This does not allow to investigate how people's judgements develop over time during presentations. This contrasts with primacy and recency theories, which suggest that some moments of the speech could be more salient than others and contribute disproportionately to the perception of the speaker's performance. To provide novel insights on this phenomenon, we present the 3MT_French dataset. It contains a set of public speaking annotations collected on a crowd-sourcing platform through a novel annotation scheme and protocol. Global evaluation, persuasiveness, perceived self-confidence of the speaker and audience engagement were annotated on different time windows (i.e., the beginning, middle or end of the presentation, or the full video). This new resource will be useful to researchers working on public speaking assessment and training. It will allow to fine-tune the analysis of presentations under a novel perspective relying on socio-cognitive theories rarely studied before in this context, such as first impressions and primacy and recency theories. An exploratory correlation analysis on the annotations provided in the dataset suggests that the early moments of a presentation have a stronger impact on the judgements.

Keywords Corpus · Public speaking · Annotation scheme · First impressions · Primacy-recency effect

1 Introduction

Public speaking constitutes a real challenge for a large part of the population: estimates indicate that 15 to 30% of the population suffers from speaking anxiety when speaking in public (Tillfors & Furmark, 2007). The automatic evaluation of public speaking performance could help in the creation of novel types of applications for training communication skills. However, it remains a complex task

Extended author information available on the last page of the article

Published online: 23 March 2024

due to its subjectivity and the challenges posed by the multi-modality of human communication.

Several works have attempted to identify the verbal and non-verbal behaviours influencing the judgements of a speaker's performance, which would be useful for creating models for the automatic assessment of public speaking trainees and for providing personalized feedback. The majority of these approaches rely on temporal-aggregate measures of the speaker's behaviours, computed from corpora collected in experimental settings (e.g., Wörtwein et al. (2015); Ramanarayanan et al. (2015)). Others focused on the possibility to assess the speaker's performance by looking at thin slices taken randomly from the overall performance (Chollet & Scherer, 2017).

In this paper, we present a new dataset, the 3MT_French dataset, aiming at facilitating the analysis of public speaking judgements, addressing some challenges of existing corpora. In particular, the dataset contains human annotations given during different moments of a presentation. This would allow for the analysis of a presentation quality under a novel perspective relying on socio-cognitive theories rarely studied before in this context, such as first impressions and primacy and recency theories.

2 Related work

Several works focused on multi-modal modelling of public speaking behaviour in different contexts, such as student presentations (e.g.,Nguyen et al. (2012)), job interviews (e.g., Naim et al. (2015); Hemamou et al. (2019)), simulations of different topic presentations (e.g.,Batrinca et al. (2013); Chen et al. (2015); Wörtwein et al. (2015); Ramanarayanan et al. (2015)), academic talks (Curtis et al., 2015), or political speech (e.g., Scherer et al. (2012)).

In this section, we review important studies in this field, discussing the datasets used for their analyses, their criteria to assess public speaking quality, and whether they took into account the temporal location of behaviours.

2.1 Existing corpora

2.1.1 Ad-hoc experimental corpora

In the context of public speaking, researchers often analyse ad-hoc datasets collected for the purposes of their study (e.g.Niebuhr and Michalsky (2018); Valls-Ratés et al. (2022)). Some of these corpora have been analysed several times. For example, Wörtwein et al. (2015) provide a multi-modal corpus collected in the context of their experimental study investigating the potential of interactive virtual audiences for public speaking training. Data from 45 speakers, each giving four presentations, was gathered to compare pre- and post-training performance in front of the Cicero virtual audience system (Chollet et al., 2014) providing different feedback. This dataset has been analysed in further studies (e.g., Chollet et al. (2021)) and integrated with additional annotations on three 10-second thin slices randomly selected from each

video (Chollet & Scherer, 2017). The particularity of this corpus is that, given the purpose of the study, the difference between pre- and post-training performance is measured. That is, it does not provide judgements for each presentation, but rather whether the performance improved or not compared to the pre-training session. In addition, the presentations are collected in an experimental setting in front of a virtual audience, that is not an ecological public speaking setting.

Another dataset, analysed in several studies (e.g., Chen et al. (2015)) contains audio, visual and Kinect data from 17 speakers, each giving four 4-5-minute presentations (both pre-prepared and improvised, without previous training and without audience). Human ratings about the presentation quality are provided, using 9 items from the Public Speaking Competence Rubric (PSCR) plus a holistic judgement. Similarly to (Wörtwein et al., 2015), this corpus is collected in an experimental setting where presentations are simulated without the presence of a real audience. In addition, it contains a relatively low number of speakers.

2.1.2 Monologues

In the context of audio-video-based job interviews, a relatively large amount of corpora were created, which usually contain monologues of candidates answering to questions from mock structured interviews, along with experts' annotations of hirability and automatically extracted audio-visual features (e.g., Naim et al. (2015); Rasipuram and Jayagopi (2016); Chen et al. (2017)). The largest one is that from Chen et al. (2017), containing 1891 monologues from 260 online participants.

In contrast to the above corpora featuring mock interviews, Nguyen et al. (2014) gathered data from 62 real job interviews, providing audio-visual features form both the interviewees and the interviewer.

Another corpus of monologues but not related to job interviews is the Persuasive Opinion Multimedia (POM) dataset (Park et al., 2014). It consists of 1000 online movie review videos. These videos are annotated for multiple speaker personality traits and high-level attributes such as confidence, credibility, entertaining, and persuasiveness.

2.1.3 Naturalistic corpora

If we focus on a context where presentations are delivered to a real audience, outside a laboratory setting, TED (Technology, Entertainment, Design) Talks[1] represent a resource with high potential. Ratings on a list of 14 adjectives (such as persuasive, inspiring, confusing) are provided. More precisely, the viewers of TED videos can annotate a talk choosing at most three of these adjectives for each talk.

A few works exist on predicting these TED Talk ratings automatically. In most cases, such works focus on transcripts, acoustic and linguistic features (e.g., Liu et al. (2017); Tanveer et al. (2019)), although others use visual features as well Sharma et al. (2018).

[1] https://www.ted.com/

Another dataset of naturalistic presentations was gathered by Curtis et al. (2015). It contains recordings of 31 academic talks given at an international conference. The particularity of this corpus is that it contains both the videos of the speaker and the audience. Audience engagement and presentation quality have been manually annotated online on 30-second segments.

2.1.4 Limitations

To summarise, several existing corpora were previously used to model public speaking behaviour. Some of them are not publicly available, for example for privacy reasons (e.g., Hemamou et al. (2019)) or because they were only released for specific challenges (e.g.,Ochoa et al. (2014)). Those created ad-hoc for specific research purposes often provide a limited amount of speakers, and are collected in an experimental setting without a real human audience. In monologues, the interaction with the audience is mostly asynchronous. In addition, most of them are collected in the context of job interviews and so the annotations are focused on hirability. TED Talks videos are a great resource but have the risk of containing mostly high-quality presentations given by expert speakers, making it difficult to investigate the behaviours related to low-quality speeches or to anxious speaking behaviour. Moreover, the videos are very long (10 min on average) and the annotation protocol is quite complex as the ratings are collected as counts instead of using more standard Likert scales.

More generally, in most of the existing corpora the annotations of the presentation quality are given *after watching the entire video*. This could limit more detailed analyses on the dynamics of the speaker's perception during the presentation.

2.2 Assessment of public speaking quality

The assessment of public speaking quality is highly subjective and depends on several interpersonal communication factors, including both verbal and non-verbal behaviours of the speaker (Baccarani & Bonfanti, 2015). This is reflected in the lack of standard evaluation criteria in most of the studies cited above, where different items were used. Nevertheless, we can notice some common categories between the evaluation rubrics used across these studies, and the tendency to ask for an additional overall assessment.

Batrinca et al. (2013) use a set of 21 typical behaviours and observable characteristics of public speaking performances such as vocal features (e.g., flow of speech, clear intonation, interrupted speech, speaks too quietly, vocal variety), body features (e.g., paces too much, gestures to emphasise, gestures to much), gaze (e.g., gazes at audience, avoids audience), as well as an overall assessment of the performance. Wörtwein et al. (2015) use a list of 10 items including eye contact, nonverbal behaviours, confidence level and an overall assessment of the performance. Chollet and Scherer (2017) reduce this list to four categories: confidence, overall performance, speech and body language. Another list of characteristics is used in the works analysing the Oral Presentation Quality Corpus provided in Ochoa et al. (2014). It includes categories related to the presentation delivery skills, such as the

structure and connection of ideas, use of voice and language, body language, eye contact and self-confidence, as well as categories related to the quality of the visual support (slides). Other studies adapt their items from the established Public Speaking Competence Rubric (PSCR) (Schreiber et al., 2012). This includes 11 items related to the speech organisation, use of language, vocal expression, non-verbal behaviour, adaptation to the audience and persuasiveness. PSCR is often completed with an overall judgement of the speaking performance, like in Chen et al. (2015) and Ramanarayanan et al. (2015). Differently from the above works, Curtis et al. (2015) simply ask annotators to rate the speaker, based on their acoustic and visual behaviour, according to the statement: "This is a good speaker who is able to capture the attention of the audience and bring the presentation to life.".

2.2.1 Common dimensions

Most of the items used to assess a presentation quality are explicitly related to the speaker's *verbal* and *non-verbal* behaviours. A few items are related to the raters' *perception* of the speaker, beyond their behaviour, and mainly concern the perceived level of *persuasiveness* and *self-confidence*.

2.3 Thin slices and temporal location of behaviours

In most of the studies cited above, the judgements about the presentation quality are given by watching the full videos, using time-aggregated features. A few others explored the temporal location of behaviours. For example, Ramanarayanan et al. (2015) focus on the dynamics of a speaker's behaviours during a presentation to predict the global quality of their presentation. Their analyses include time-series features, computed through histograms of co-occurrence of different features such as head pose, eyes gaze and facial expressions. These features, used independently or combined with time-aggregated ones, have been found to be useful for prediction of different public speaking ratings. Chollet and Scherer (2017) investigate the use of *thin slices* of behaviours (Ambady & Rosenthal, 1992) for assessing public speaking performance. They consider three slices of 10-second randomly selected from the full video of each speaker. The ratings given from the thin slices are highly correlated with those of the full videos, and show that it is possible to predict ratings of a presentation quality using audio-visual features extracted from the thin slices.

This latter study, as well as a few others like (Nguyen & Gatica-Perez, 2015), demonstrate that it is possible to predict public speaking quality from thin slices randomly selected from a presentation, but they do not focus on the temporal location of these slices. Previous work shows that the moments that are most important in a speech are the beginning and the end. For example, the *primacy and recency effect* (Ebbinghaus, 1913) is exploited by politicians as a persuasive strategy in their speech (e.g., Hongwei (2020)). A similar effect is also found in the context of job interviews. The analyses in (Hemamou et al., 2021) on peaks of attention slices (of a duration between 0.5s and 3.3s) during asynchronous job interviews show that these slices are systematically different from random slices. They occur more often

at the beginning and end of a response, and are better than random slices at predicting hirability. On the other hand, *first impressions* theory (Ambady & Skowronski, 2008) argues that perceivers form an impression of others at the earliest instants of an interaction (the earliest instants of a speech in our case), and that this first impression is hard to modify subsequently. If this theory applies to our context, we should find a significant impact of the speakers' behaviour at the *beginning* of their speech, and what happens during the rest of the speech should be less discriminative about their performance. Finally, it could also be that what is important for a speaker is to maintain the listener's attention during the speech. In this case, their behaviour at the *middle* of the speech should be more informative about their performance.

3 Motivation and contributions

Being motivated to face most of the limits discussed in the previous section, and to facilitate the investigation of how people's judgements develop during a speech, we present the 3MT_French dataset of public speaking presentations. Compared to previous work, this dataset allows for a novel perspective for the analysis of a presentation quality, relying on first impressions and primacy and recency theories (see Sect. 2.3).

With this work, we aim at providing two types of contributions. On one hand, the 3MT_French dataset with its particular properties:

- A relatively large amount (248) of naturalistic presentations given in front of a real audience;
- The speakers are not necessarily experts in public speaking, which means that the quality of the presentations is highly heterogeneous;
- The presentations all have the same duration (180 s) and follow a similar structure;
- Information about speakers who won audience and/or jury prize is included.

On the other hand, we also provide the following methodological contributions that can be useful for other domains:

- A novel annotation scheme is proposed, which aims at providing a quick way to rate the quality of a presentation, considering the dimensions in common between other existing schemes;
- The annotations are collected for both the entire video and at different time windows.

4 The 3-minute thesis competition

The 3-minute Thesis competition was originally conceived by the University of Queensland in 2008 and is now held in over 900 universities across more than 85 countries worldwide. It allows PhD students to present their research topic, in

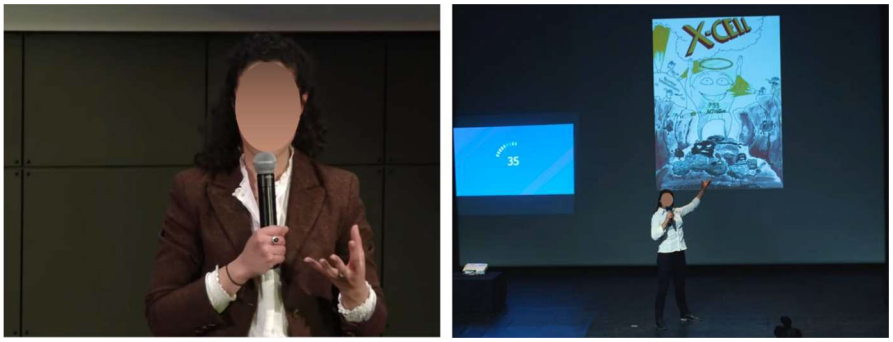

Fig. 1 Two screenshots from the presentations available in the 3MT_French dataset. The first focuses on the speaker's gestures and facial expressions; the second one includes the speaker's full body and the slide used as a support

simple terms, to a non-expert and diverse audience. Each student must make a clear, concise and convincing presentation of their research project in no more than 180 s. One single slide can be used to support the presentation.

The concept was taken up in 2012 in Quebec and extended to all French-speaking countries. In particular, the French edition of the competition, called "Ma thèse en 180 ses", has been held since 2014.[2] The competition begins with the selection of representatives from each university, which may be open to the non-scientific audience. Regional rounds are organised between candidates from different universities, followed by national semi-finals and finals and an international final including other French-speaking countries.

4.1 The 3MT_French dataset

In our 3MT_French dataset, we focus on the French edition of the competition, held in 2019, which was the last year where the presentations took place in presence in front of the audience, without any physical restriction (the 2020 and 2021 editions were held partially or totally virtually, due to the pandemic). We selected the presentations from the regional rounds, the first phase whose videos were published online (upon participants' agreement). Since they still are at the beginning of the competition, we can find a high variety of presentations, and not only high-quality ones as may be the case for the national final. The 3MT_French dataset contains videos from 248 presentations, 135 of female and 113 of male speakers, annotated on several time windows (see Sect. 5.2). Their videos are publicly available on YouTube (the URLs, start and end time stamps are provided in the dataset). Two screenshots are shown in Fig. 1.

[2] https://mt180.fr/

4.2 Jury and audience prizes

At each regional round, a 1st prize is awarded by a jury composed of experts in public speaking and scientific mediation, and another prize is awarded from the votes of the audience. According to the regional rules, a 2nd and a 3rd jury prize may be awarded. We can consider the jury prizes as an objective judgement of the presentation, as the jury members followed a list of specific criteria, while the audience prize is a more subjective judgement, as the audience was just asked to vote for their favourite presentation, without any instructions.

The dataset is unbalanced with respect to the number of winners, with 58 out of 248 (23%) of presentations winning a prize (9 speakers won both a jury and the audience prize). On the other hand, there is no effect of gender on winning a prize: neither for the jury prize ($\chi^2(1)$ with Yate's correction$= 0, p = 1$) nor for the audience prize ($\chi^2(1)$ with Yate's correction$= 0.12, p = 0.73$).

5 Annotations of presentations' quality

One of the goals related to the development of the 3MT_French dataset is to propose a novel annotation scheme useful to assess the quality of a presentation. Several annotation schemes, described in Sect. 2, were used in previous work. Similarly, the jury prizes assigned to the participants of the French 3MT competition follow an evaluation grid (see Sect. 4.1). Each of them (previous schemes and 3MT grid) focuses on different criteria, but we can identify some common dimensions, related to both the speaker's behaviour and the perception of the speaker. With *speaker's behaviour* we intend both verbal and non-verbal cues, including vocal features, speech content, gestures, gaze, etc (more details in Sect. 2.2). With *perception of the speaker*, we refer to criteria beyond the actual behaviours, such as the level of persuasiveness and perceived self-confidence of the speaker. Most of the previous annotation schemes also include the assessment of the overall performance.

As a trade-off between using a comprehensive but time-expensive and a quick but limited annotation scheme, we decided to focus on the subjective perception of the speaker, without explicitly asking about the use of behaviours. The latter information can be automatically extracted without necessarily requiring human annotation, while subjective dimensions like persuasiveness do not. The potential limitations related to these particular variables are discussed in Sect. 6.4.

5.1 Annotation scheme

The proposed annotation scheme aims at providing a quick way to rate the perception of a public speaking quality, considering several dimensions in common between the existing schemes. In previous work, the most frequent items concern the perceived level of persuasiveness and self-confidence of the speaker. In addition, we consider the perception of the audience engagement during the presentation.

This aspect has been investigated in a few works, for example Curtis et al. (2015) analysed both videos of the speaker and the audience and found that it is possible to predict levels of audience engagement based on the speaker's verbal and non-verbal behaviours. In our case, the videos of the presentations do not allow for assessing the audience engagement from the audience behaviour itself. What we ask to the raters is to provide their perception of the audience engagement according to the speaker's behaviour. This task turned out to be potentially ambiguous, as we discuss in Sect. 6.4.

Below we detail all the items of the proposed annotation scheme.

5.1.1 Introduction

Before the raters watched a presentation and completed the annotation task, we highlighted that the task was only for French-speaking participants. It was important for us that raters understood the content of the speech, so that their annotations could be used to investigate the role of textual features in public speaking perception. Compared to audio and visual features, only a few studies focused on textual features (e.g., Larrimore et al. (2011), Yang et al. (2020)). The 3MT_French dataset will allow to develop research in this direction. Then, we introduced the 3MT competition, and asked the raters to answer to questions taking into account the speaker's behaviour. We highlighted the fact that they should watch the video entirely, and we specified that it may be cut at the beginning and/or at the end.

5.1.2 Global evaluation

For the first dimension of the annotation scheme, we asked the raters to give their global evaluation of the presentation without focusing on specific criteria or dimensions. We just provided some benchmarks on a 100-point Likert scale. Given the absence of specific criteria, the 100-point scale was used instead of more standard 5 or 7 points, on one hand to allow for more nuanced answers, and on the other hand to facilitate the task by referring to a familiar scoring system (reminding grading and percentages). Thus, relying on how we introduced the context of the competition, the question we asked is the following (note that this is the English translation of the original French question): "*Give an overall score for the presentation, on a scale from 1 to 100, where: 1= the presentation is not at all acceptable for this type of competition; 25 = the presentation is quite poor and could not win at any level of competition; 50= the quality of the presentation does not allow me to say if it could win or not; 75= the presentation is good enough to win some phases of the competition, but not the final; 100= the presentation is perfect and could definitely win the competition.*"

5.1.3 Persuasiveness

We asked raters to annotate their perception of the level of persuasiveness of the speaker, according to the definition given in the PSCR (item 11: "Constructs an effectual persuasive message with credible evidence and sound reasoning."). Thus,

the English translation of the question we asked is: "*In your opinion, on a scale from 1=not at all to 5=very much, how persuasive is the person in the video, i.e., do they effectively craft a convincing message? Is their reasoning rigorous?*"

5.1.4 Perceived self-confidence

Speaking self-confidence has been identified as being the same construct as self-perceived communicative competence (SPCC) (Yu et al., 2011; Lockley, 2013). SPCC concerns how competent people feel they are in a variety of communication contexts and with a variety of types of receivers (McCroskey & McCroskey, 1988). Applying this concept to a third-party observation, the perceived level of confidence of a speaker can be rated as their competence to effectively accomplish their preferred outcomes in ways perceived as appropriate to the context and by the communication (Morreale et al., 2015; Spitzberg, 2000).

The English translation of the question in our annotation scheme is the following: "*In your opinion, on a scale of 1=not at all to 5=very much, how competent is the person in the video, i.e. are they an expert in their field? How effectively do they convey their message in a contextually appropriate way?*"

5.1.5 Audience engagement

Engagement is a complex process, for which a large variety of definitions exist across different domains. We focus here on the concept of engagement as defined by researchers in human-computer interaction (for a review, see Oertel et al. (2020)), since they usually refer to the same phenomena occurring between humans. In particular, Peters et al. (2009) distinguish between the attentional and emotional components of engagement. The former can be defined as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another" (Sidner & Dzikovska, 2002). Emotional engagement, on the other hand, involves empathy and could be defined as "the fostering of emotional involvement intending to create a coherent cognitive and emotional experience which results in empathic relations [...]"(Scherer, 2000). The two components interleave, as the attention is driven by emotions. Thus, the English translation of the question we asked to rate the perception of audience engagement is the following: "*On a scale from 1=not at all to 5=very much, to what extent does the audience stay attentive and maintain an emotional connection with the speaker?*" As mentioned above, the audience was rarely visible during the presentation, thus the idea was that the rater mainly inferred engagement from the speaker's behaviour rather than the actual audience's behaviour.

5.1.6 Control question

As a necessary condition to validate the task, we included a control question to check that the raters actually understood French. This was important to ensure that the speech content was taken into account during the annotation. Participants had to find a specific verb in the instructions and to conjugate it at a specific person and

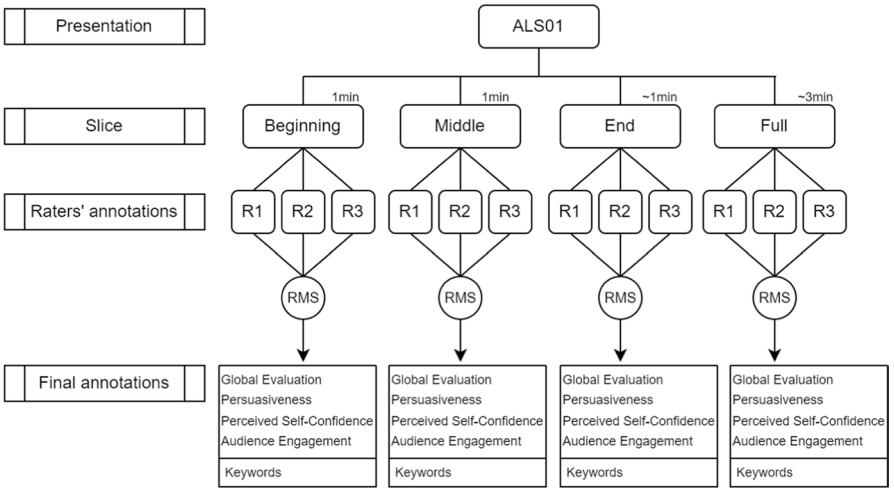

Fig. 2 The terms used in our annotation protocol. Each *presentation* includes four *slices* (i.e., beginning, middle, end and full), each of them annotated by three raters R. The *final annotations* for each slice contains the root mean square (RMS) of the three *rater's annotations*, for each of the measured variables (i.e., global evaluation, persuasiveness, perceived self-confidence, audience engagement), as well as the keywords related to the topic of the presentation

tense. The answer was automatically checked and only the right one allowed for submitting the task.

5.1.7 Keywords

In addition to the annotation scheme related to the perception of the presentation quality, we also asked the raters to provide one or more keywords related to the topic of the presentation. The original reason for it was to check if the raters watched the videos carefully and to double-check their French comprehension (in addition to the control question, see above). The keywords have been manually validated by the authors of this work. This brings additional content to the 3MT_French dataset, that could be exploited for research purposes. For example, it could be investigated whether the agreement between the raters in the choice of the keywords is reflected in their agreement about the presentation quality.

5.2 Annotation protocol

To facilitate the understanding of the protocol followed to collect the annotations, we use a specific terminology, as depicted in Fig. 2. The term *presentation* indicates the speech of each speaker, which is represented by a unique label; for example, ALS01 is the presentation of the first speaker (in alphabetical order) of the Alsace regional round. We use the term *slice* to indicate the annotated videos, including three 1-minute slices (beginning, middle and end) and a full slice (3 min) for each

presentation. The *beginning* slice includes the first minute of the presentation, the *middle* slice includes the second minute, while the *end* slice includes the third minute. Note that the end slice could last a bit less than 60 s, depending on the actual length of the presentation. Similarly, the *full* slice could last a bit less than 3 min.

Each slice was annotated by three raters, for a total of 248 presentations $*$ 4 slices $*$ 3 raters = 2976 annotations. A *rater's annotation* includes the ratings of the variables detailed in Sect. 5.1. The annotations were collected through the Amazon Mechanical Turk platform (Buhrmester et al., 2016), giving a reward of 0.40 and 1 for each 1-minute and 3-minute full slices, respectively. The average duration to annotate a 1-minute slice was of around 3 min, and around 5 min for the 3-minute full slice.

Each rater was free to annotate as many slices as they wanted, but only annotated each presentation once. That is, taking the ALS01 presentation as an example, it is split into four slices to annotate (i.e., ALS01-beginning, ALS01-middle, ALS01-end and ALS01-full). A rater could be assigned to only one of these four slices. The answers from raters who did not watch the videos entirely were discarded and replaced by new annotations. The same procedure was applied to annotations whose keywords were off-topic or not in French.

The set of the *final annotations* provided in the 3MT_French dataset contains, for each slice of each presentation, the root mean scores (RMS) of persuasiveness, perceived self-confidence, audience engagement and global evaluation, as well as the set of the keywords provided by the three raters. Already existing information about jury and audience prizes and the speaker's gender is also included. The 3MT_French dataset is available on Zenodo.[3]

6 Descriptive analyses

In this section, we report some descriptive analyses related to the annotations collected through the Amazon Mechanical Turk platform. As stated above, the new dataset and the novel annotation scheme presented in the paper are proposed as a novel perspective for the analysis of a public speaking judgement, relying on first impressions and primacy and recency theories. Accordingly, the analyses presented in this section focus on the correlations between the measured variables (i.e., persuasiveness, perceived self-confidence, global evaluation), the observed slices (i.e., beginning, middle, end or full) and the judgements given during the competition (i.e., jury and audience prizes).

6.1 Scores

As described in Sect. 5.2, each video was annotated by a different random set of three raters. This condition does not allow for computing the consistency of the

[3] https://zenodo.org/record/7603511

Table 1 ICC scores for each variable and for each slice

	Persuasiveness	Perceived Self-confidence	Audience Engagement	Global Evaluation
Beginning	0.22	0.17	0.12	0.17
Middle	0.38	0.25	0.03	0.36
End	0.19	0.08	0.18	0.19
Full	0.11	0.14	0.04	0.12

scores within the raters, but only inter-rater *absolute agreement*, i.e., the extent to which the different raters tend to give exactly the same score when rating the same video (Tinsley & Weiss, 1975). It may occur that raters rely on different internal scales, as has been found when assessing affective content (Metallinou & Narayanan, 2013; Yang & Chen, 2010).

The intraclass correlation coefficient (ICC) (Bartko, 1966) is the most suitable for our protocol as it can be used for ordinal data and takes into account the fact that each slice is rated by a different set of randomly chosen raters (raters are considered as random effects). In particular, we computed a one-way random, average score ICC (McGraw & Wong, 1996) for each variable and slice. The ICC values are reported in Table 1. Each line corresponds to one slice (the 1-minute slices beginning, middle and end, and the 3-minute full video). Each column corresponds to the annotated dimensions described in Sect. 5.1.

The low values of agreement between the raters are not surprising, indeed it is a common issue when performing subjective annotations in the context of social computing studies (Salminen et al., 2018) or when rating emotion databases (Siegert et al., 2014), especially when using crowdsourcing (Karpinska et al., 2021). Inspired from suggestions in (Siegert et al., 2014) and (Karpinska et al., 2021), we limited the risk of high variance by providing context information and carefully checking the French-speaking requirement and the time spent to complete the annotation task. Nevertheless, we cannot exclude that our annotation scheme may contribute to the low agreement (see Sect. 6.4).

When looking at the ICC scores across the slices, it seems to be a tendency to lower agreement when annotating the full video, and highest scores for the middle slice (except for audience engagement). This could indicate that it is often more difficult to give or agree on a judgement when considering a complete presentation, while assessing specific, local moments is more straightforward.

Interestingly, the agreement is generally lower for audience engagement. This could be due to the fact that, differently from the other variables, the raters were asked to judge this variable without having continuous information about the audience's reactions. The relatively higher agreement for the end slice would support this hypothesis, as it could be related to the potential presence of applause at the very end of the slice. Anyway, the sparsely reactions of the audience make audience engagement not fully exploitable for the 3MT_French dataset, but we

Table 2 Correlations (Pearson's r) between annotations of persuasiveness (P), perceived self-confidence (SC) and global evaluation (GE), for the beginning, middle, end and full slices. All $p < 0.05$

Beginning			Middle			End			Full		
	SC	GE		SC	GE		SC	GE		SC	GE
P	0.8	0.76	P	0.83	0.81	P	0.83	0.79	P	0.83	0.79
SC	–	0.62	SC	–	0.77	SC	–	0.74	SC	–	0.74

Table 3 Correlations (Pearson's r) of the annotations across the slices (b: beginning, m: middle, e: end, f: full) and the audience (a) and jury (j) prizes for perceived self-confidence, persuasiveness and global evaluation. All $p < 0.05$, ns: $p > 0.05$

Perceived self-confidence						Persuasiveness						Global Evaluation					
	m	e	f	a	j		m	e	f	a	j		m	e	f	a	j
b	0.16	0.17	ns	0.16	ns	b	0.17	0.24	0.14	0.17	ns	b	0.24	0.19	0.16	0.15	ns
m	–	0.24	0.23	ns	ns	m	–	0.25	0.24	ns	ns	m	–	0.22	0.32	ns	ns
e	–	–	0.16	ns	ns	e	–	–	0.15	ns	ns	e	–	–	0.14	ns	ns
f	–	–	–	ns	ns	f	–	–	–	ns	ns	f	–	–	–	ns	ns

believe that it would be an interesting variable to annotate for other corpora containing audience's videos. This variable is not further analysed here.

In order to handle the general low agreement between annotators, the consensus between the annotators is then built by computing the root mean square (RMS), as made by Dinkar et al. (2020).

6.2 Correlations between variables

Table 2 reports the Pearson's r values for each correlation between the annotated variables at each slice. All the ratings are highly positively correlated. This halo effect has already been reported in previous studies focusing on similar dimensions, like in (Chollet & Scherer, 2017).

In particular, we can observe that persuasiveness and perceived self-confidence have the highest values of correlation, with $r \geq 0.8$ no matter the slice. Another interesting observation is that the halo effect of these two variables with the global evaluation scores is slightly amplified (i.e., higher correlation scores) when rating the middle slice. This may suggest a stronger impact of perceived self-confidence and persuasiveness as influencing the global evaluation of a speech during that part of the presentation.

6.3 Correlations between slices and with the prizes

Table 3 reports the Pearson's r values of the correlations between the different slices for each variable. The last two columns, i.e., a: audience and j: jury, are binary

variables where 1 indicates that a presentation won the audience (or jury, respecitvely) prize, 0 otherwise.

In general, the ratings are slightly correlated between slices for each dimension. This somehow reassures us about the consistency of the annotations across the different moments of the presentations.

Previous work already shows that it is possible to replace full videos with thin slices (Chollet & Scherer, 2017). The correlation scores between the ratings given after watching the full video and those related to the other slices may inform about what parts of the speech are more representative of the full video. The results tend towards the middle slice, which shows the highest correlation with the full video, for all the variables.

When looking at the correlations between the collected annotations and the judgements given during the competition, we can see that they are correlated with the audience prize variable only when considering the beginning slice, while no correlations are found neither for the other slices nor with the jury prize. We could speculate that the crowdsourced raters and the audience are closer because they are both non-experts compared to the jury. In addition, the presence of a correlation only at the beginning of the presentations could suggest an impact of first impressions on public speaking judgements, but could also indicate that other variables than the ones investigated in this paper were considered during the competition. This result seems to indicate that the final judgement is made at a certain point of the presentation, relatively early. In a future work, it would be interesting to investigate how to better determine when this moment occurs.

6.4 Discussion and limitations

The descriptive analyses highlight some interesting characteristics of the annotations, which should be taken into account by researchers willing to use the 3MT_French dataset for their analyses.

First of all, the inter-rater agreement is relatively low. Even if this is a common issue for crowdsourced data, we can identify some factors which potentially contributed to these low scores, some of which may rely to the proposed annotation scheme. The high-level variables we focused on are very subjective and their ratings could vary according to individual characteristics like personality and culture. However, we share the thoughts of Leonardelli et al. (2021), that is, "disagreement should be seen as a signal and not as noise". One could specifically focus on the videos with lowest inter-rater agreement to investigate what behavioural cues are more difficult to judge. It should be noticed that subjectivity may not be the only explanation to the low inter-rater agreement, in particular in the case of audience engagement annotations. This variable was difficult to evaluate without continuous information about the audience's reactions, as shown by the relatively higher agreement for the end slice, where the presence of applause could have facilitated the annotation. In addition, the question itself could be ambiguous. The original purpose was that the participants evaluated the audience engagement according to the speaker's behaviour. The question may have had different interpretations, for example, raters may have

rated their own engagement instead. As future work, it would be interesting to ask raters to annotate their own engagement to investigate the impact of sharing (or not) the same physical space with the speaker on the performance perception.

Another potential limitation of the ratings is that the annotation task was restricted to French speakers. As mentioned in Sect. 5.1.1, we wanted to provide ratings obtained while understanding the content of the speech. This makes the 3MT_French dataset suitable for investigating the role of textual features (both lexical and semantic) in public speaking assessment, which is still a little explored topic, in particular for French. This restriction may have two drawbacks. On one hand, it does not allow to distinguish the impact of non-verbal behaviour and the speech content, since we do not have annotations made by raters who did not understand French. On the other hand, we did not check for the culture or nationality of participants but only their French comprehension, thus there could be a variability in the annotations due to culture that we cannot control.

As future work, we are aiming to collect additional annotations on the 3MT_French dataset, in particular to open to non-French participants and to investigate how culture affects the perception of the speaker. It would also be interesting to ask participants which aspects influenced mostly they ratings (e.g., body behaviour, speech, voice, etc.). This information could show whether inter-rater disagreement is related to different behaviours taken into account during the annotation process.

7 Conclusion

We presented the 3MT_French dataset, a new corpus for the analysis of public speaking quality. It contains the presentations of PhD students participating in the French edition of 3-minute Thesis competition. The particularity of the dataset is that the information about the jury and audience prizes awarded during the competition has been integrated with a set of ratings collected online through a novel annotation scheme and protocol. Global evaluation, persuasiveness, perceived self-confidence and audience engagement have been annotated at different time windows (i.e., the beginning, middle or end of the presentation, or the full video). Keywords related to the topic of each video are also available.

This new resource would interest several researchers working on public speaking assessment and training, as well as it will allow for perceptive studies, both under a behavioural and linguistic point of view. It will allow for investigating whether a speaker's behaviours have a different impact on the observers' perception of their performance according to *when* these behaviours are realised during the speech. The automatic assessment of a speaker's performance could benefit from this information by assigning different weights to segments of behaviour according to their relative position in the speech. In addition, a training system could be more efficient by focusing on improving the speaker's behaviour during the most important moments of their performance.

The second contribution of this paper is the development of a new annotation scheme that could be used on other public speaking datasets in addition to the 3MT_French one. Its purpose is to provide a quicker and reliable alternative to the large

amount of existing schemes, by focusing on the *perception* of the performance and considering the common dimensions previously used by other authors.

Acknowledgements The authors would like to thank the CNRS and CPU for permitting the analysis of the videos, as well as the reviewers for their useful suggestions. This work was partially funded by the Carnot institutes TSN and M.I.N.E.S. under the Inter-Carnot contract 200000830 AI4SoftSkills and the ANR-21-CE33-0016-02 REVITALISE project.

Declarations

Ethical approval The videos annotated in the 3TM_French dataset are publicly available on YouTube and their rights belong to CNRS and CPU, who gave us their approval to analyse the videos for research purposes. One potential ethical concern related to the dataset would be to use the data to automatically judge the quality of a speech. This is not the purpose of our work, since we are interested in understanding how people form their judgements and to develop training tools to improve public speaking skills. As in all research activities involving human beings, gender differences may exist. Potential gender bias need to be addressed by the researchers interested in using the 3MT_French dataset as an integral part of their analyses to ensure the highest level of scientific quality. The gender of the speaker is provided, while no information about the gender of the annotators was collected through the Amazon Mechanical Turk platform.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256.

Ambady, N., & Skowronski, J. J. (2008). *First impressions*. Guilford Press.

Baccarani, C., & Bonfanti, A. (2015). *Effective public speaking: A conceptual framework in the corporate-communication field*. Corporate Communications: An International Journal.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*(1), 3–11.

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2013). Cicero-towards a multimodal virtual audience platform for public speaking training. International workshop on intelligent virtual agents (pp. 116–128).

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*(1), 3–5.

Chen, L., Leong, C.W., Feng, G., Lee, C.M., & Somasundaran, S. (2015). Utilizing multimodal cues to automatically evaluate public speaking performance. In: 2015 International Conference on Affective Computing and Intelligent Interaction (acii) (pp. 394–400).

Chen, L., Zhao, R., Leong, C.W., Lehman, B., Feng, G., & Hoque, M.E. (2017). Automated video interview judgment on a large-sized corpus collected online. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (acii) (pp. 504–509).

Chollet, M., Marsella, S., & Scherer, S. (2021). Training public speaking with virtual social interactions: Effectiveness of real-time feedback and delayed feedback. *Journal on Multimodal User Interfaces*. https://doi.org/10.1007/s12193-021-00371-1

Chollet, M., & Scherer, S. (2017). Assessing public speaking ability from thin slices of behavior. In: 2017 12th ieee International Conference on Automatic Face & Gesture Recognition (fg 2017) (pp. 310–316).

Chollet, M., Sratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2014). An interactive virtual audience platform for public speaking training. Pro-ceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (pp. 1657–1658).

Curtis, K., Jones, G.J., & Campbell, N. (2015). Effects of good speaking techniques on audience engagement. In: Proceedings of the 2015 acm on International Conference on Multimodal Interaction (pp. 35–42).

Dinkar, T., Colombo, P., Labeau, M., & Clavel, C. (2020). The importance of fillers for text representations of speech transcripts. arXiv preprint arXiv:2009.11340 .

Ebbinghaus, H. (1913). Memory: a contribution to experimental psychology. 1885. Teachers College, Columbia University.

Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J.-C., & Clavel, C. (2019). Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In: Proceedings of the aaai Conference on Artificial Intelligence (Vol. 33, pp. 573–581).

Hemamou, L., Guillon, A., Martin, J.-C., & Clavel, C. (2021). Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact recruiter's decision. IEEE Transactions on Affective Computing.

Hongwei, Z., et al. (2020). Analysis of the persuasive methods in Barack Obama's speeches from the social psychology's perspectives. *The Fron- tiers of Society, Science and Technology, 2*(10), 11–16.

Karpinska, M., Akoury, N., & Iyyer, M. (2021). The perils of using mechanical turk to evaluate open-ended text generation. arXiv preprint arXiv:2109.06835 .

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D. M., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research, 39*, 19–37.

Leonardelli, E., Menini, S., Palmero Aprosio, A., Guerini, M., & Tonelli, S. (2021,November). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10528–10539). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlpmain.822 10.18653/v1/2021.emnlp-main.822

Liu, Z., Xu, A., Zhang, M., Mahmud, J., & Sinha, V. (2017). Fostering user engagement: Rhetorical devices for applause generation learnt from ted talks. In: Proceedings of the International aaai Conference on Web and Social Media (Vol. 11).

Lockley, T., et al. (2013). Exploring self-perceived communication competence in foreign language learning. *Studies in Second Language Learning and Teaching, 3*(2), 187–212.

McCroskey, J.C., & McCroskey, L.L. (1988). Self-report as an approach to measuring communication competence.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30.

Metallinou, A., & Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In: 2013 10th ieee International Conference and Workshops on Automatic Face and Gesture Recognition (fg) (pp. 1–8).

Morreale, S., Staley, C., Stavrositu, C., & Krakowiak, M. (2015). Firstyear college students' attitudes toward communication technologies and their perceptions of communication competence in the 21st century. *Communication Education, 64*(1), 107–131.

Naim, I., Tanveer,M.I., Gildea, D., & Hoque,M.E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In: 2015 11th ieee International Conference and Workshops on Automatic Face and Gesture Recognition (fg) (Vol. 1, pp. 1–6).

Nguyen, Chen, W., & Rauterberg, M. (2012). Online feedback system for public speakers. 2012 ieee symposium on e-learning, e-management and e-services (pp. 1–5).

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of Hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia, 16*(4), 1018–1031.

Nguyen, L.S., & Gatica-Perez, D. (2015). I would hire you in a minute: Thin slices of nonverbal behavior in job interviews. In: Proceedings of the 2015 acm on International Conference on Multimodal Interaction (pp. 51–58).

Niebuhr, O., & Michalsky, J. (2018). Virtual reality simulations as a new tool for practicing presentations and refining public-speaking skills. In: 9th International Conference on Speech Prosody 2018 (pp. 309–313).

Ochoa, X.,Worsley,M., Chiluiza, K., & Luz, S. (2014). Mla'14: Third multimodal learning analytics workshop and grand challenges. In: Proceedings of the 16th International Conference on Multimodal Interaction (pp. 531–532).

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., & Peters, C. (2020). Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI, 7*, 92.

Park, S., Shim, H.S., Chatterjee, M., Sagae, K., & Morency, L.-P. (2014). Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In: Proceedings of the 16th International Conference on Multimodal Interaction (pp. 50–57).

Peters, C., Castellano, G., & De Freitas, S. (2009). An exploration of user engagement in hci. In: Proceedings of the International Workshop on Affective-Qware Virtual Agents and Social Robots (pp. 1–3).

Ramanarayanan, V., Leong, C.W., Chen, L., Feng, G., & Suendermann-Oeft, D. (2015). Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In: Proceedings of the 2015 acm on International Conference on Multimodal Interaction (pp. 23–30).

Rasipuram, S., & Jayagopi, D.B. (2016). Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: a systematic study. In: Proceedings of the 18th acm International Conference on Multimodal Interaction (pp. 370–377).

Salminen, J.O., Al-Merekhi, H.A., Dey, P., & Jansen, B.J. (2018). Inter-rater agreement for social computing studies. In: 2018 Fifth International Confer-ence on Social Networks Analysis, Management and Security (snams) (pp. 80–87).

Scherer, S., Layher, G., Kane, J., & Neumann, H., & Campbell, N. (2012). *An audiovisual political speech analysis incorporating eye-tracking and perception data* (pp. 1114–1120). LREC.

Scherer, K. (2000). Emotion. introduction to social psychology: A european perspective. m. hewstone and w. stroebe. Oxford.

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competence rubric. *Communication Education, 61*(3), 205–233.

Sharma, R., Guha, T., & Sharma, G. (2018). Multichannel attention network for analyzing visual behavior in public speaking. In: 2018 ieee Winter Conference on Applications of Computer Vision (wacv) (pp. 476–484).

Sidner, C.L., & Dzikovska, M. (2002). Human-robot interaction: Engagement between humans and robots for hosting activities. In: Proceedings. fourth ieee International Conference on Multimodal Interfaces (pp. 123–128).

Siegert, I., Böck, R., & Wendemuth, A. (2014). Inter-rater reliability for emotion annotation in human-computer interaction: Comparison and methodological improvements. *Journal on Multimodal User Interfaces, 8*(1), 17–28.

Spitzberg, B. H. (2000). What is good communication? *JACA: Journal of the Association for Communication Administration, 29*(1), 103–19.

Tanveer, M.I., Hassan, M.K., Gildea, D., & Hoque, M.E. (2019). Predicting ted talk ratings from language and prosody. arXiv preprint arXiv:1906.03940 .

Tillfors, M., & Furmark, T. (2007). Social Phobia in Swedish University Students: Prevalence, subgroups and avoidant behavior. *Social Psychiatry and Psychiatric Epidemiology, 42*(1), 79–86.

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*(4), 358.

Valls-Ratés, Ï., Niebuhr, O., & Prieto, P. (2022). Unguided virtual-reality training can enhance the oral presentation skills of high-school students. *Frontiers in Communication, 7*, 196.

Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelhagen, R., & Scherer, S. (2015). Multimodal public speaking performance assessment. In: Proceedings of the 2015 acm on International Conference on Multimodal Interaction (pp. 43–50).

Yang, Y.-H., & Chen, H. H. (2010). Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(4), 762–774.

Yang, Z., Huynh, J., Tabata, R., Cestero, N., & Aharoni, T., & Hirschberg, J. (2020). What makes a speaker charismatic? Producing and perceiving charismatic speech. *Speech Prosody, 2020*, 685–689.

Yu, H., Li, H., & Gou, X. (2011). The personality-based variables and their correlations underlying willingness to communicate. *Asian Social Science, 7*(3), 253.

Authors and Affiliations

Beatrice Biancardi[1] · Mathieu Chollet[2] · Chloé Clavel[3,4]

✉ Mathieu Chollet
mathieu.chollet@glasgow.ac.uk

Beatrice Biancardi
bbiancardi@cesi.fr

Chloé Clavel
chloe.clavel@telecom-paris.fr

[1] CESI LINEACT, Nanterre, France

[2] School of Computing Science, University of Glasgow, Glasgow, UK

[3] LTCI, Télécom Paris, Palaiseau, France

[4] INRIA Paris, INRIA, Paris, France