2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition

Assessing Public Speaking Ability From Thin Slices of Behavior

Mathieu Chollet and Stefan Scherer
Institute for Creative Technologies, University of Southern California, Playa Vista, CA, USA

***Abstract*—An important aspect of public speaking is delivery, which consists of the appropriate use of non-verbal cues to strengthen the message. Recent works have successfully predicted ratings of public speaking delivery aspects using the entire presentations of speakers. However, in other contexts, such as the assessment of personality or the prediction of job interview outcomes, it has been shown that thin slices, brief excerpts of behavior, provide enough information for raters to make accurate predictions. In this paper, we consider the use of thin slices for predicting ratings of public speaking behavior. We use a publicly available corpus of public speaking presentations and obtain ratings of full videos and thin slices. We first study how thin slices ratings are related to full video ratings. Then, we use automatic audio-visual feature extraction methods and machine learning algorithms to create models for predicting public speaking ratings, and evaluate these models for predicting thin slices ratings and full videos ratings.**

I. INTRODUCTION

When considering what contributes to a great speech, there is no doubt that the careful selection of arguments, the use of a compelling vocabulary and the organization of the speech into a coherent structure all matter a great deal for its overall quality. However, the delivery of the speech, *i.e.* the appropriate mobilization of vocal, facial and bodily cues to support the speech is another crucial aspect that has been recognized as essential in rhetoric by scholars since classical antiquity: as the Roman orator Cicero mentioned, "*In the matter of delivery which we are now considering, the face is next in importance to the voice.*" [16].

Public speaking is an important skill, as it can influence our careers and personal lives. Unfortunately, it is also one of the most commonly reported social phobias [5], [14]. Therefore, it is important to consider ways to support the training and assessment of public speaking, both for individuals suffering from public speaking anxiety and for non-anxious individuals simply willing to improve their ability. Recently, several interactive systems have been introduced for public speaking training with real-time feedback [3], [9], [12], [28], [30], [31]. Immediate feedback has been shown to be beneficial for improving public speaking behavior [17], and evaluations of those real-time feedback systems have indeed shown that they produced beneficial outcomes. However, they usually focus on providing feedback on a limited set of non-verbal behaviors at at time, *e.g.* pause fillers or gaze direction [9], volume and speaking rate [30] or body energy and speaking rate [12]. To that end, they rely on some form of user input, received either through way of a Wizard-of-Oz (*i.e.* a confederate produces the system's input by monitoring the trainee and pressing appropriate keys when relevant behaviors are produced) [9] or through automatic processing of the relevant audio-visual signals [12], [30]. None of these public speaking systems have so far used automatically assessed measures of the user's holistic multimodal public speaking behavior.

Nevertheless, recent attempts at creating models for automatically scoring public speaking presentations from audio-visual recordings have proved to be successful [32], [24], [11]. For training those models, full videos of public speaking performances were used by their respective authors. In the context of an application providing real-time feedback on your immediate public speaking behavior, it could be that models trained on complete presentations would not accurately report the quality of the trainee's recent behavior. Instead, real-time assessment of public speaking behavior would have to rely on relatively short segments of public speaking behavior. Such small segments of behavior have been referred to as "thin slices" [1], [7]. Thin slices have been the subject of extensive research; it was shown that exposure to short recordings of an individual can be enough to judge complex individual constructs such as personality or intelligence [1], [7], or even social outcomes such as job interview hiring decisions [22], negotiation outcomes [10] or performance in collaborative tasks [18].

In this paper, we propose to investigate the use of thin slices of behavior to produce automatic assessments of public speaking from multimodal behavior. We investigate the following research questions:

- Q1 Are ratings obtained from thin slices as reliable as ratings obtained from entire presentations, and what is their relationship?
- Q2 Is it possible to automatically assess public speaking performance with audio-visual features extracted from thin slices? Specifically:
  - (a) Can we predict thin slices ratings?
  - (b) Can we predict full videos ratings?
  - (c) Can we predict public speaking anxiety?

In the next section, we present related work on automatic multimodal behavior assessment and on the use of thin slices. In section III, we introduce the dataset of public speaking presentations. We then present our results and discuss their implications in section IV.

II. RELATED WORK

A. Multimodal assessment of public speaking performance

Several researchers have realized experiments to automatically predict ratings of politicians [26], [29] and debaters' [6] speeches. Researchers found that vocal variety, as measured by fundamental frequency (f_0) range and maximal f_0 of

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.45

focused words are correlated with perceptual ratings of a good speaker or debater [29], [6]. Further, manual annotations of disfluencies were identified to be negatively correlated with a positive rating. In [27], the acoustic feature set used in [29] was complemented by measures of pause timings and measures of tense voice qualities. The study shows that tense voice quality and reduced pause timings were correlated with overall good speaking performances. Further, the authors investigated visual cues, in particular motion energy, for the assessment of the speakers' performances. They found that motion energy is positively correlated with a positive perception of speakers.

Wörtwein *et al.* investigated unimodal and multimodal assessment of public speaking presentations [32]. They used a greedy feature selection method to train regression ensemble trees to predict various aspects of public speaking performance, and found that combining multimodal features usually works better than using audio or visual features only. For instance, the features selected by their model for overall performance assessment, achieving a correlation of $\rho = 0.745$, were related to vocal variety and to facial expressions. They achieved similarly high performance for predicting public speaking anxiety [33], their models reaching a correlation of $\rho = 0.825$.

In [24], the authors compared the relative utility of multimodal features of different natures, *i.e.* time-aggregated features (*e.g.* mean or standard deviation of a feature over a complete presentation) and time-series features (*e.g.* histograms of co-occurence for expressions from different modalities). The advantage of such features time-aggregated ones is that they allow to capture temporal behavior patterns, for instance how often a certain facial expression follows another facial expression. The authors found that time-series based features were useful for prediction of different public speaking ratings, either on their own or in combination with time-aggregated features.

Curtis *et al.* proposed to predict not only a public speaker's presentation ratings but the level of audience engagement using multimodal features [11]. They used a dataset of conference scientific presentations containing audience views and presenter views, that they annotated for speaker performance and audience engagement using ordinal scales. They analyzed correlations between multimodal features and these annotations and found that speaker ratings had the highest correlations with speaker motion and articulation rate, whilst audience engagement correlated the highest with speech intensity features. They then built ordinal class classifiers for these two ratings, achieving accuracies of 73.1% for classifying speaker ratings and 70.3% for audience engagement.

B. Thin slices

Thin slices have a long history in psychology research, and have been studied in a large number of contexts, from the prediction of individual traits such as personality. In the remainder of this section, we report a number of works that have focused on automatic assessments using multimodal features from thin slices. In [2], Ambady and Rosenthal published one of the first meta-analyses on the topic. A notable result was that the duration of thin slices did not seem to affect judgment accuracy for social outcomes or individual traits (*e.g.* existence of deception, physician proficiency and patient satisfaction); in particular, thin slices shorter than 30s did not differ significantly from longer slices. Carney *et al.* examined the reliability of slices of various lengths (5s, 20s, *etc*, up to 300s) for the recognition of affect, personality and intelligence [7]. Contrary to Ambady *et al.* [2], they found that increases in slice duration increased accuracy: for instance, while slices of 5s were sufficient for judging some traits (intelligence, negative affects, neuroticism, conscientiousness), 20s slices were significantly more accurate for others (positive affects, agreeableness, openness). On the behavior level, recent results by Murphy *et al.* [20] combining data from 4 different studies on the accuracy of thin slices for assessing overall behavior (*e.g.* percentage of speaking time, number of nods or gestures) found that slices longer than 30s may be required to reflect the actual behavior tendencies of a subject.

In [4], Batrinca *et al.* investigated the automatic assessment of personality using multimodal features from self-presentation videos lasting between 30 and 120 seconds. Subjects were recorded with a frontal camera and a microphone. Visual features were obtained with manual annotation, while audio features were extracted automatically. The authors cast their problem as a binary classification problem between low and high values of the 5 dimensions of the Big-5 personality model [25]. Their method allowed to reach classification accuracies ranging from 65% (agreeableness dimension) to 76% (emotional stability dimension).

Curhan and Pentland presentend an approach to predict negotiation outcomes from thin slices of behaviors [10]. The negotiation in question consisted of a negotiation classroom exercise, where one subject acts as an employee which must negotiate their compensation package to their company's vice-president, played by the other subject. They used the 5 first minutes of every interaction, from which they extracted a varied set of features: conversational features, such as speaking time or overlapping speech, prosodic features, such as pitch or volume, and interactional features such as vocal mimicry. They then created regression models, which accounted up to 30% of the total variance of the data. One interesting result they found is that for the two roles, *i.e.* employee or vice-president, the negotiation success was predicted by different features.

Lepri *et al.* attempted to predict individual performances of subjects participating in the Mission Survival corpus [18]. In this corpus, groups of 4 participants have the objective to assign priorities to objects that would be the most helpful to survive in a harsh environment after a plane crash. Using audio-visual and interactional features extracted from the first minute of the interaction, they tried to classify the performance of individual members between low, medium, and high performances. The best performance they achieved was an above-chance accuracy of 0.41, when considering only features from the classified individual (no features from other participants).

Nguyen *et al.*'s realized an experiment to try to predict the hirability of applications using multimodal cues from thin slices of behavior [22]. They used slices of variable durations (from a few seconds to more than 2 minutes) in order to match the questions and answers structure of the job interviews. They found that multimodal models trained to predict the interview outcome on thin slices were as accurate as human resource professionals using the same thin slices.

III. MULTIMODAL CORPUS

In order to investigate our research questions, we used a publicly available multimodal corpus of public speaking presentations [8].

A. Original corpus

The corpus we used in our study consists of audio-visual recordings of 45 subjects performing 4 successive presentations in front of the Cicero public speaking virtual audience system [9]. Out of the 45 subjects, 27 were male and 18 were female, and the mean age was 37 years (σ = 12.7). Presentations lasted 4 minutes in average. In the first (*pre-test*) and last (*post-test*) presentations, the virtual audience was passive while in the two intermediate presentations (*training*) the audience could be interactive, with differences in feedback type depending on the training condition. Additionally, the *pre-* and *post-*presentations shared the same topic, *i.e.* the city in which the study was realized (Los Angeles, USA), and the 2 *training* presentations shared another topic, the presentation of a beauty product. In this paper, we only use the first (*pre-test*) and last (*post-test* presentations as they share the same topics and as the virtual audience behavior could have influenced the behavior of the subjects in the other presentations.

The corpus includes audio, visual, and Kinect data for each presentation as well as questionnaires for each subject, including the 'Big Five Inventory' personality questionnaire [25] and the 'Personal Report of Confidence as a Speaker (PRCS)' questionnaire [23], used to estimate public speaking anxiety [15]. Additionally, the authors made available a collection of extracted audio-visual features, including voice quality measures extracted with the COVAREP speech analysis toolbox [13], facial expressions of emotions extracted with the FACET tool[1], and features of body activity obtained from the Kinect. We provide a summary of the available audio-visual features in table I, and the full description of these features can be found in [8].

Finally, the authors included evaluations of various public speaking performance aspects realized by experts recruited from a local Toastmasters association and by laypeople recruited on the Amazon Mechanical Turk[2] crowdsourcing website. Whilst the experts annotations are interesting, those evaluated the *difference* between performance between the *pre-* and *post-test* videos, meaning that those do not give us information as the actual performance of the subjects in each video. Therefore, we decided to use the MTurk crowdsourced measures for the full videos, consisting of ratings of 5 different raters for each video. The annotated performance aspects were derived from the literature and discussion with experts and are the following:

- Eye Contact
- Body Posture
- Flow of Speech
- Gesture Usage
- Intonation
- Confidence Level
- Stage Usage
- Avoids pause fillers
- Presentation Structure
- Overall Performance

[1]http://www.emotient.com/products

[2]https://www.mturk.com/

Identifier	Description	Source
Visual features		
GestureK	Upper body activity	Kinect
OrientationK	Orientation wrt. the audience	Kinect
EmotionF	Facial expression of emotions: *e.g.* Anger, Contempt, Joy...	Facet
GazeO	Eye contact	OKAO
SmileO	Amount of smiling	OKAO
Audio features		
f0	Fundamental frequency (pitch)	Matlab
Energy	Loudness of audio signal	Matlab
VUV	Voiced/unvoiced ratio (speaking ratio)	COVAREP
NAQ, QOQ, H1H2, PSP, MDQ, RD	Voice quality features	COVAREP
MFCC0-3	Mel-frequency Cepstrum Cepstrum Coefficients 0-3	COVAREP
KF1-3	First 3 KARMA-filtered formants	KARMA

TABLE I: Listing of the audio-visual features used. For visual features, the features' means are used on the whole slice. For audio features, we use the features' mean and standard deviations.

B. Thin slices dataset

For our experiment, we also created a dataset of thin slices from the presentations in [8]. There are several important considerations when investigating thin slices of behavior, two of which being the *duration* of the slices and the *location* of the slices within the complete videos [20]. As we presented in section II-B, a general take-away from the literature is that while longer slice lengths tend to increase the accuracy of ratings, the difference is usually not significant from a shorter slice. Considering that we investigate automatic assessment in the context of potential real-time feedback for public speaking training, we wanted to reduce the length of the slice to reflect the very recent behavior of a public speaker. Indeed, immediate feedback has been shown to wield considerable power in certain contexts, in particular public speaking [17]. As previous results show that thin slices shorter than 30s do not yield less accurate judgments of social outcomes [2], and in an effort to make the slices short enough for speakers to remember their immediate behavior in a training setting, we settled on a slice length of 10s. For choosing the slice locations within the complete presentations, we decided to simply apply a random selection (making sure there was no overlap between them). Indeed, in the context of a real-time assessment application, we would have to analyze behavior from the beginning of a training session until its end. Finally, we decided to extract 3 different slices per video in order

Full videos			
Pause fillers	0.278	Body posture	0.236
Confidence	0.402	Eye contact	0.377
Flow of speech	0.369	Gesture usage	0.519
Overall performance	0.428	Presentation structure	0.313
Speech intonation	0.32	Stage usage	0.272
Thin slices			
Confidence	0.391	Speech	0.29
Performance	0.369	Non-verbal	0.35

TABLE II: Krippendorf's α scores for full videos ratings and thin slices ratings.

to ensure the amount of ratings to be obtained wouldn't be too large: were we to extract all thin slices (without overlap) from the corpus, then about 2200 videos would have to be annotated.

We then proceeded to obtain annotations of public speaking performance for the thin slices dataset. We initially planned to use the same measures for the full videos and for the thin slices. However, we realized that some of these categories were too broad or inapplicable for short videos, *e.g.* the presentation structure. Therefore, we decided to instead ask turkers for a limited set of 4 constructs. We kept the same *Confidence* and *Overall Performance* aspects as for the full videos, and re-grouped the vocal and bodily aspects into two larger ones, *Speech* and *Body Language*. We collected ratings of 4 different turkers for every pre- and post-training video slice, making sure that every turker could only rate one slice for a single presenter.

C. Inter-rater agreement

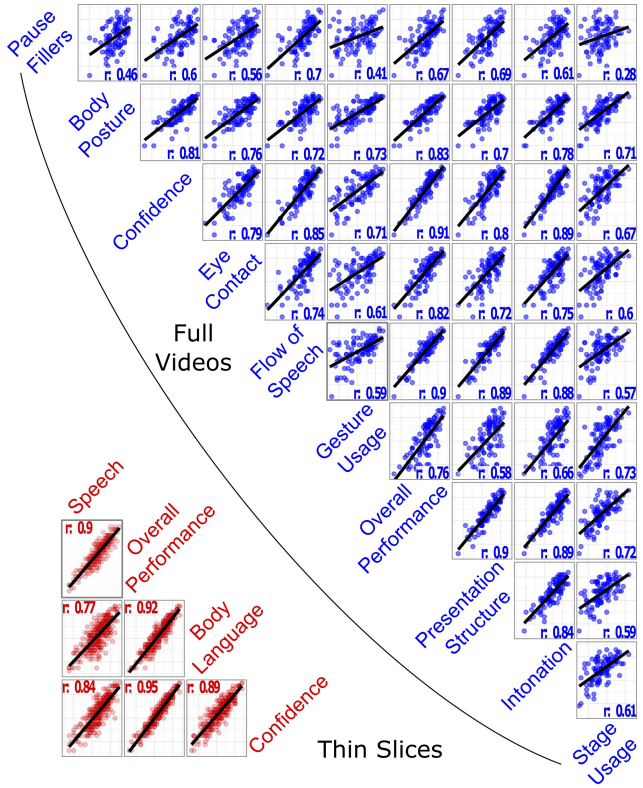
We computed measures of inter-rater agreement, namely Krippendorff's α, in the case of full videos ratings and thin slices ratings in order to investigate whether raters tend to disagree more in one of the two cases. Results are shown in table II for both full video annotations and thin slices.

As can be observed, α values fall in general within the $0.25 - 0.5$ range. While those agreement measure can seem poor, they are in line with other researchers' findings [24], [21], showing that there may be important differences in appreciation of public speaking performance. Additionally, we notice that agreements do not seem to differ strongly between ratings of full videos and ratings of thin slices.

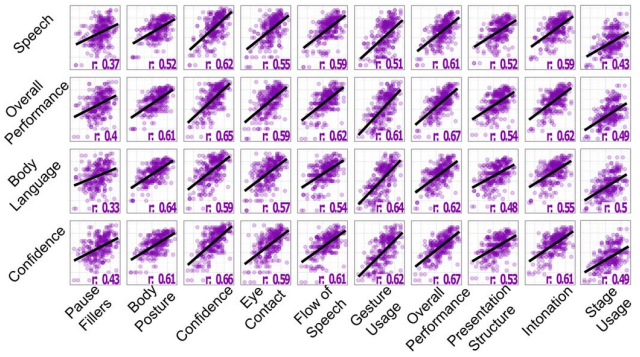
IV. RESULTS

A. Comparison of thin slices ratings and full videos ratings

Our first research question was directed at investigating whether ratings of public speaking behavior from thin slices of behavior are related to ratings made from full performances. To investigate this, we computed correlations between the different variables annotated by turkers. In Figure 1a, we plot



(a) Correlations between the 10 full videos annotated variables (top right, blue) and correlations between the 4 thin slices annotated variables (bottom left, red) videos.



(b) Correlations between the full videos annotations and the thin slices annotations.

Fig. 1: Correlation graphs for the various rated aspects of public speaking performance.

the correlation graphs for the full video variables and thin slices variables separately. We observe that there are very strong correlations between all the annotated variables: even with a Bonferroni correction, all 91 correlations (45 intra-correlations for full videos, 6 intra-correlations for thin slices, 40 inter-correlations) are significant ($p < 0.0005$).

For full video annotations, the weakest correlation is to be between *stage usage* and *pause filler avoidance* ($\rho = 0.28$),

and the highest between *confidence* and overall performance ($\rho = 0.91$). For thin slices annotations, the lowest is between *speech* and *body language* ($\rho = 0.77$) while the highest correlation is between *overall performance* and *confidence* ($\rho = 0.95$). The inter-correlations between thin slices variables and full videos variables are shown in Figure 1b. The lowest correlation is found between thin slices ratings of *body language* and full video ratings of *pause fillers* ($\rho = 0.33$) and the highest correlations between thin slices ratings of *overall performance* and *confidence* and full video ratings of *overall performance* ($\rho = 0.67$).

From those results, we can already notice that ratings made on thin slices of public speaking behavior are highly correlated with ratings of full videos. Therefore, we suppose that it may be possible to predict the performance ratings of full videos based on audio-visual features extracted from thin slices only.

B. Prediction of public speaking performance

In this section, we then try to predict ratings of public speaking performance using thin slices of behavior. In a first part, we only analyze the thin slices videos: we extract audio-visual features from those videos and then try to predict the ratings of that video. In a second part, we try to predict the scores of full public speaking presentations, based on the features of thin slices only. Finally, we try to predict public speaking anxiety ratings derived from the PRCS questionnaires filled out by participants before the experiments.

For those three parts, we adopt the same methodology as Wörtwein *et al.*: we compare prediction performance of regression tree ensembles using visual features only, audio features only and audio-visual features together. The features of those separate sets are presented in table I. We apply a greedy sequential feature selection method in order to identify the most relevant features and to reduce the amount of features in the models. The training procedure is the following: A test set comprised of 30% of the data is held out. On the training set, the feature selection is performed using a leave-one-speaker out cross-validation strategy. Once the features are selected and the tree ensemble are trained, their prediction performance is evaluated on the held out testing set.

We present our results in table III. In the upper part of the table, ratings of thin slices are predicted using visual, audio, and multimodal features from those slices (*Q2a*). In the middle part, ratings of full videos of public speaking performance are predicted from thin slices features (*Q2b*). Finally, in the last line we display results for predicting public speaking anxiety scores, obtained from the PRCS questionnaires [23], using features from thin slices (*Q2c*).

Our results show that it is possible to predict ratings of public speaking performance using audiovisual features extracted from 10s long thin slices randomly sampled in a public speaking presentation. Indeed, our models achieve high correlations with ground truth ratings on a majority of annotated categories, such as $\rho = 0.624$ for *flow of speech* during the whole interaction, or $\rho = 0.598$ for the *overall performance* on a specific thin slice. The hardest aspect to automatically derive from thin slices features is the *avoidance of pause fillers* ($\rho = 0.349$), which may be unsurprising as pause fillers may not always be present during the extracted video segments. Conversely, the best correlation is attained for the *flow of speech* ($\rho = 0.624$).

V. DISCUSSION

We started this paper by defining two main research questions revolving around thin slices of public speaking presentation.

A. Relationship of thin slices ratings and full video ratings

In our first question *Q1*, we asked whether annotator ratings of public speaking performance realized on full videos and on thin slices would be related. We investigated this in section IV-A by computing correlations between thin slices ratings and full video ratings. All the considered variables show very high inter-correlations, the lowest being between the rating of the avoidance of pause fillers on full videos and the rating of speech on thin slices ($\rho = 0.33$), going as high as $\rho = 0.67$ for the ratings of overall performance for both cases.

Another interesting result from that investigation is that there are very high correlations between all the various aspects annotated by raters in both the full video case, and the thin slices case. This suggests that the overall performances of public speakers seem to affect the ratings of all the separate aspects of their behaviors.

B. Automatic assessment of public speaking from thin slices

Secondly, in order to answer *Q2*, we attempted in section IV-B to create models for predicting the different annotated aspects from audio-visual features automatically extracted from thin slices. For most considered aspects, our models were able to reach correlations between predicted scores and original ratings in the $0.4 - 0.6$ range. This is an encouraging result, as one of our objectives is to be able to provide feedback to public speaking trainees on their immediate performance.

For the case of predicting the thin slices ratings (*Q2a*), our best models achieve correlations between $\rho = 0.461$ for the rating of *Speech* and $\rho = 0.598$ for *Performance*. This validates that immediate public speaking behavior can be automatically assessed, and that such assessments could be used to provide feedback in real-time. Unsurprisingly, the visual features reach higher performance for predicting the *Body Language* rating; the *GestureK* is actually the most informative features for predicting this rating using audio-visual features. However, using visual features seems to generally lead to poorer results. Both audio-only and audio-visual feature sets seem to perform well in general.

Moving on to the prediction of full videos (*Q2b*), an interesting observation is that there does not seem to be very large differences in prediction performance between ratings of full videos (*Q2b*) and ratings of thin slices (*Q2a*). This seems to indicate that the thin slices approach is also valuable for predicting the overall public speaking ability of an individual. Similarly to the thin slices case, the audio-visual feature set

	Visual	Acoustic	Acoustic+Visual
Thin Slices Confidence	0.025 (1.106) GestureK	**0.571 (0.871)** μPSP, μMFCC1, μMFCC2	0.469 (0.936) μf0, μH1H2, μPSP
Thin Slices Body Language	0.275 (0.931) GestureK, OrientedK	0.459 (0.923) μf0, μPSP, μMFCC0	**0.495 (0.898)** GestureK, μMFCC1 , μMFCC2
Thin Slices Performance	0.080 (0.999) GestureK	0.450 (0.850) μVUV, μMFCC1, μKF3	**0.598 (0.727)** μf0, μVUV, μMFCC0
Thin Slices Speech	-0.037 (0.966) GestureK	0.358 (0.866) μMDQ, μMFCC0, μMFCC1	**0.461 (0.790)** μNAQ, μRD, μMFCC0
Pause Fillers	-0.101 (0.979) GestureK	0.265 (0.973) μRD, μMFCC0, μMFCC1	**0.349 (0.902)** μVUV, μPSP, μRD
Body Posture	0.260 (0.697) GestureK	0.194 (0.818) μH1H2, μPSP, μKF3	**0.400 (0.723)** GestureK, μVUV, μRD
Confidence	0.122 (1.059) GestureK	0.430 (0.992) μVUV, μMFCC0, σf0	**0.522 (0.905)** μVUV, μMFCC1, μKF3
Eye Contact	0.052 (0.989) GestureK	0.246 (1.011) μQOQ, μMFCC0, μMFCC1	**0.318 (0.857)** GestureK, μQOQ, μMFCC0
Flow Of Speech	0.321 (0.775) GestureK	0.596 (0.772) μVUV, μMFCC0, σf0	**0.624 (0.684)** μMFCC0, μKF1, μKF2
Gesture Usage	0.184 (1.293) GestureK	**0.494 (1.094)** μVUV, μMFCC0, μNAQ	0.316 (1.280) μMFCC1, μKF2, σKF3
Overall Performance	-0.006 (0.898) GestureK	0.500 (0.811) μVUV, μNAQ, μMFCC0	**0.527 (0.783)** μVUV, μMFCC0, σMFCC0
Presentation Structure	0.097 (0.908) GestureK	0.486 (0.884) μVUV, μPSP, μMDQ	**0.510 (0.702)** μVUV, μMFCC2, σKF2
Speech Intonation	0.117 (0.813) GestureK	**0.614 (0.730)** μPSP, μRD, μKF1	0.533 (0.719) μNAQ, μMDQ, μMFCC1
Stage Usage	**0.533 (0.781)** GestureK	0.280 (0.878) μQOQ, μPSP, μMFCC1	0.479 (0.921) GestureK, μMFCC1, μMFCC2
Anxiety	0.145 (0.266) GestureK, OrientedK	**0.405 (0.230)** μf0, μRD, μKF1	0.199 (0.257) μQOQ, σH1H2, σPSP

TABLE III: Results of our experiment to predict public speaking ability aspects. For every cell, the format is the following: $\frac{correlation\ (Mean\ absolute\ error)}{selected\ features}$. For each rating category, the feature set performing best is shown in bold, and the three best features are displayed using their identifiers defined in table I. For audio features, we denote the mean and standard deviation functionals with μ and σ.

dominates with 7 out of 10 highest scores. Additionally, it seems that visual features were also not very informative there, with the exception of the prediction of *stage usage*, which is best predicted by the *GestureK* measure of gestural activity ($\rho = 0.533$). The *flow of speech* rating seems to also be well predicted by visual features, and the *gesture usage* rating is best predicted by audio features: these exceptions may be explained by the intimate relationship of speech and gesture [19]. Finally, while our models reach adequate performance, in particular for holistic features such as *performance* ($\rho = 0.527$) or *confidence* ($\rho = 0.522$), their performance is still lower than what previous approaches reached using features extracted from the full videos, and not just thin slices [32]. This is in line with earlier findings on thin slices durations, which indicate that increasing slice durations allow to reach more accurate judgments [7], [20].

Finally, we investigated the prediction of public speaking anxiety, obtained using PRCS questionnaires [23], from features extracted from thin slices. In this case, highest performance is reached using audio features only ($\rho = 0.405$). Similarly to the prediction of full video performance ratings, we reach lower performance with thin slices than approaches using full video features [33], however in this case the difference is more striking. This indicates that public speaking anxiety may require stimuli of longer length to be adequately recognized.

VI. CONCLUSION

In this paper, we investigated the use of thin slices of behavior for the evaluation of public speaking performance. Using a publicly available multimodal corpus of presentations, we compared ratings of public speaking behavior realized on thin slices and ratings realized on full videos. We then proceeded to build prediction models using automatically extracted audio-visual features from the thin slices videos. In general, we found that thin slices are a very relevant approach for evaluating public speaking behavior. User ratings of full videos and thin slices are very related, and prediction models trained on thin slices audiovisual features can achieve high correlations with original annotations.

Thin slices have been used in many domains and our results contribute new evidence to the fact that they may indeed carry a lot of relevant information. In future work, we intend to use our prediction models in interactive public speaking training

applications, in order to give immediate and relevant feedback to trainees regarding their immediate public speaking behavior. In particular, we will study whether a thin slices approach can be applied to determine, in public speaking training sessions , which segments of a speaker's behavior were the best and the worst, in order to allow the trainees to review their training presentation efficiently and to reflect on what aspects of their behavior they should improve.

VII. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant No. IIS-1421330 and U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government, and no official endorsement should be inferred.

REFERENCES

[1] N. Ambady, F. J. Bernieri, and J. A. Richeson. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in experimental social psychology*, 32:201–271, 2000.

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

[3] R. Barmaki and C. E. Hughes. Providing real-time feedback for student teachers in a virtual rehearsal environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 531–537. ACM, 2015.

[4] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 255–262. ACM, 2011.

[5] G. D. Bodie. A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication Education*, 59(1):70–105, 2010.

[6] M. Brilman and S. Scherer. A multimodal predictive model of successful debaters or how i learned to sway votes. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 149–158. ACM, 2015.

[7] D. R. Carney, C. R. Colvin, and J. A. Hall. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072, 2007.

[8] M. Chollet, T. Wörtwein, L.-P. Morency, and S. Scherer. A multimodal corpus for the assessment of public speaking ability and anxiety. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 488–495. ELRA, 2016.

[9] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer. Exploring feedback strategies to improve public speaking: An interactive virtual audience framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1143–1154. ACM, 2015.

[10] J. R. Curhan and A. Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802, 2007.

[11] K. Curtis, G. J. Jones, and N. Campbell. Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 35–42. ACM, 2015.

[12] I. Damian, C. S. S. Tan, T. Baur, J. Schöning, K. Luyten, and E. André. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 565–574. ACM, 2015.

[13] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP - A collaborative voice analysis repository for speech technologies. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 960–964. IEEE, 2014.

[14] T. Furmark, M. Tillfors, H. Stattin, L. Ekselius, and M. Fredrikson. Social phobia subtypes in the general population revealed by cluster analysis. *Psychological Medicine*, 30(6):1335–1344, 2000.

[15] J. Hook, C. Smith, and D. Valentiner. A short-form of the personal report of confidence as a speaker. *Personality and Individual Differences*, 44(6):1306–1313, 2008.

[16] G. A. Kennedy. *The Art of Rhetoric in the Roman World: A History of Rhetoric*. Princeton University Press, 1972.

[17] P. E. King, M. J. Young, and R. R. Behnke. Public speaking performance improvement as a function of information processing in immediate and delayed feedback interventions. *Communication Education*, 49(4):365–374, 2000.

[18] B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi. Automatic prediction of individual performance from thin slices of social behavior. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 733–736. ACM, 2009.

[19] D. McNeill. *Gesture and thought*. University of Chicago Press, 2008.

[20] N. A. Murphy, J. A. Hall, M. Schmid Mast, M. A. Ruben, D. Frauendorfer, D. Blanch-Hartigan, D. L. Roter, and L. Nguyen. Reliability and validity of nonverbal thin slices in social interactions. *Personality and Social Psychology Bulletin*, 41(2):199–213, 2015.

[21] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Proceedings of the 11th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6. IEEE, 2015.

[22] L. S. Nguyen and D. Gatica-Perez. 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*, pages 51–58. ACM, 2015.

[23] G. Paul. *Insight vs. Desensitization in Psychotherapy: An Experiment in Anxiety Reduction*. Stanford University Press, 1966.

[24] V. Ramanarayanan, C. W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft. 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*, pages 23–30. ACM, 2015.

[25] B. Rammstedt and O. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of research in Personality*, 41(1):203–212, 2007.

[26] A. Rosenberg and J. Hirschberg. Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of Interspeech*, pages 513–516, 2005.

[27] S. Scherer, G. Layher, J. Kane, H. Neumann, and N. Campbell. An audiovisual political speech analysis incorporating eye-tracking and perception data. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1114–1120. ELRA, 2012.

[28] J. Schneider, D. Börner, P. van Rosmalen, and M. Specht. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 539–546. ACM, 2015.

[29] E. Strangert and J. Gustafson. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *Proceedings of Interspeech*, pages 1688–1691, 2008.

[30] M. Tanveer, E. Lin, and M. E. Hoque. Rhema: A real-time in-situ intelligent interface to help people with public speaking,. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 286–295. ACM, 2015.

[31] M. I. Tanveer, R. Zhao, K. Chen, Z. Tiet, and M. E. Hoque. Automanner: An automated interface for making public speakers aware of their mannerisms. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 385–396. ACM, 2016.

[32] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer. Multimodal public speaking performance assessment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 43–50. ACM, 2015.

[33] T. Wörtwein, L.-P. Morency, and S. Scherer. Automatic assessment and analysis of public speaking anxiety: A virtual audience case study. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction*, pages 187–193. IEEE, 2015.