Virtual Debate Coach Design: Assessing Multimodal Argumentation Performance

Volha Petukhova
Spoken Language Systems Group, Saarland University
Saarbrücken, Germany
v.petukhova@lsv.uni-saarland.de

Tobias Mayer
Computational Linguistics, Saarland University
Saarbrücken, Germany
s9tomaye@stud.uni-saarland.de

Andrei Malchanau
Spoken Language Systems Group, Saarland University
Saarbrücken, Germany
andrei.malchanau@lsv.uni-saarland.de

Harry Bunt
Tilburg Center for Communication and Cognition
Tilburg University, The Netherlands
harry.bunt@uvt.nl

ABSTRACT

This paper discusses the design and evaluation of a coaching system used to train young politicians to apply appropriate multimodal rhetoric devices to improve their debate skills. The presented study is carried out to develop debate performance assessment methods and interaction models underlying a Virtual Debate Coach (VDC) application. We identify a number of criteria associated with three questions: (1) how convincing is a debater's argumentation; (2) how well are debate arguments structured; and (3) how well is an argument delivered. We collected and analysed multimodal data of trainees' debate behaviour, and contrasted it with that of skilled professional debaters. Observational, correlation and machine learning experiments were performed to identify multimodal correlates of convincing debate performance and link them to experts' assessments. A rich set of prosodic, motion, linguistic and structural features was considered for the system to operate on. The VDC system was positively evaluated in a trainee-based setting.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**; **Interaction techniques**; • **Applied computing → Interactive learning environments**;

KEYWORDS

Multimodal behaviour tracking, natural debate argumentation, virtual coach

ACM Reference Format:
Volha Petukhova, Tobias Mayer, Andrei Malchanau, and Harry Bunt. 2017. Virtual Debate Coach Design: Assessing Multimodal Argumentation Performance. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3136755.3136775

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136775

1 INTRODUCTION

We currently see transformative developments in novel devices and sensing technologies that get more and more interconnected and seamlessly integrated in everyday human activities. Multimodal conversational interfaces enable users to interact with their devices, appliances and other systems in an intuitive and natural manner. Multimodal dialogue is not only the most social and natural form of interaction, but has been proven to have positive effects when incorporated in human learning and medical treatment [22, 51, 74]. It has been also shown that 'digital immersion' can enhance learning [15, 30, 50]. Multi-sensory approaches not only reinforce learning, but also personalize the assessment process and engage learners.

The current state of technology enables fairly fine grained and inexpensive tracking of visible body movements and facial expressions (Intel®RealSense™, 3D Kinect, Tobii Glasses) and measuring various biometrical signals (Blood Volume Pulse and NeXus EXG sensors). While exhaustive real-time monitoring seems unrealistic, certain multimodal markers that trigger and guide the interaction and presentation of information may be defined. Progress has been made in multimodal behaviour modelling, with advances in social signal processing and affective computing [68]. The identification of multimodal markers and their relation to psycho-physiological assessments is however still under development. The main goal of the presented study is to provide theoretical framework, methodological insights and evaluation metrics for feature selection criteria that may help explain perceptive regularities and provide a set of operational (recognizable and measurable) indicators to assess multimodal human performance. The application discussed in this paper, the Virtual Debate Coach, is an interactive tutoring system designed for the training of debate skills in political contexts.

A debate is a communication process in which participants argue for or against a certain position proposed for the dispute. Whereas the argumentative elements of debating have received ample attention as a means to enhance learning [16], learning relevant aspects of debating has received less attention.

The training of debate skills typically involves ad-hoc face-to-face classroom debates. The debater's skills proficiency level is often judged on three criteria: (1) argument organization, (2) argument content, and (3) argument delivery. Argument content and organization have received considerable attention of philosophers, logicians and linguists [37, 62, 69]. Based on the conversational nature and

social construction of arguments, we propose a discourse-based model for the analysis and evaluation of debate argument structure (Section 4). Argument content evaluation methodology translates human judgments into viable and easy to compute argument quality metrics (Section 5). To assess argument delivery aspects related to persuasive debate style, a range of linguistic, paralinguistic and body language features of professional debaters are analysed (Section 6). Section 7 presents the design, module- and trainee-based evaluation of the Virtual Debate Coach. We wrap up the paper by summarising our findings and outline future research.

2 RELATED WORK

2.1 Arguments and Argumentation in Debate

An argument is defined as consisting of a statement that can be supported by evidence. A statement (*claim*) is an assertion that deserves attention. There may be a conclusion which presents a result, which can be derived from certain evidence (*premises*).

Previous work in argumentation theory and artificial intelligence was largely based on designing and applying argumentation schemes, see [17, 62, 69]. Toulmin (1958) proposed a scheme with six functional roles to describe the structure of an argument. Based on evidence (*data*) and a generalisation (*warrant*), which is possibly implicit and defeasible, a *conclusion* is derived. The conclusion can be *qualified*, e.g. by strengthening the inferential link between data and conclusion. A *rebuttal* specifies exceptional conditions that undermine this inference. A warrant can be supported by *backing*, e.g. reason, justification or motivation.

Recently, argumentation mining techniques have been applied to natural arguments analysis, see the survey in [42]. Independent of the approach, most researchers seem to agree on the theoretical skeleton of logical and pragmatic aspects - the connection between subject and predicate on a logical propositional level and the inter-propositional relations on the pragmatic level. Translating Toulmin's general argumentation scheme into a structure of debate arguments, we have premises for a claim (main statement, **A***rgument*) that can be of **R***eason* and **E***vidence* types, and a claim that may be summarised or re-stated in a conclusion, often referred as an ARE structuring technique, see [43].

Another commonly used technique to support a claim with evidence is called *chunking* [25]. Here, debaters generalise from a claim (*chunking up*), provide a specific example (*chunking down*) or draw analogies (*chunking sideways*).

Debaters are trained to follow rules imposed by the above mentioned structures, respect domain conventions and best practices.[1]

2.2 Training Argumentation Skills

Few existing argumentation training systems work for spoken discourse. For example, Ashley et al. [2] use transcripts of arguments produced in the US Supreme Court as a basis for training hypothetical reasoning drawing the similarities with the legal case in question. The trainee is shown an argument transcript and asked to build an argumentation structure graph following Toulmin's scheme. The system detects trainee's contextual and structural weaknesses, and provides feedback. Trainees do not formulate their own arguments but use pre-defined phrases, or are offered the option to substitute special legal formulations with semantically equivalent ones.

There are web-based argumentation training systems available, e.g. DebateGraph[2] and TruthMapping[3]. The former provides a platform to prevent opinion manipulation, marking inconsistent arguments in online discussions. The system represents arguments as graphs spotting unsupported premises and giving the user the possibility to rebut or support arguments. TruthMapping facilitates collaborative learning through argumentation. Arguments are also represented as graphs, different standpoints and their evidences are visualized to the learners encouraging them to address those.

In our scenario, debaters exchange 'natural' arguments, i.e. they are not constrained in the use of communicative means, and they also may exploit 'extra-rational' characteristics of their audience, taking into account emotions and affective factors. It is important, therefore, not only to understand the underlying structure of natural arguments explaining certain regularities but also to evaluate means and strategies used by debaters to deliver convincing performance.

In current educational design practice there is a growing interest in using whole-tasks models that aim to assist students in integrating knowledge, skills and attitudes into coherent wholes, to facilitate learning transfer [65]. Characteristics of a 'skilled professional debate performance' are defined in terms of coaching goals related to (1) argument organization, (2) argument content, and (3) argument delivery, see [66] for 'conducting debate' skills hierarchy.

2.3 Multimodal Properties of Convincing Debate

Debates, in particular political debates, constitute a large portion of public speeches. Skilled professional debaters give the impression that they truly believe what they say, know how to catch and keep the audience attention, express authority, confidence, respect and friendliness. People generally associate certain speech, personality and interaction features with what they think is a 'good public speaker' [57]. Debaters make a number of choices from a wide range of rhetorical, lexical, syntactic, pragmatic and prosodic devices to deliver strong persuasive speech. They often use *intensifiers*, i.e. individual words or phrases that are syntactically, tonally or rhythmically marked, *parallelisms* (word or phrase repetitions for information density reduction and emphasis, e.g. well-known 'Lists of Three' [4]), and *meta-discursive acts*[4] to relate speaker to audience, to maintain topic-comment structure, etc. [4, 38, 61]. Prosodic and acoustic strategies in speech may be decisive in conveying an opinion in a political debate [7]. Clear articulation, sufficient voice volume level, and well adjusted tempo are strongly associated with professional public speaking. Pitch range, voice and speaking rate variations are perceived as expressions of enthusiasm, engagement, commitment and charisma, see also [49]. Mispronounced words, frequent hesitations, restarts and self-corrections negatively influence the perceived speaker confidence and may jeopardize speaker credibility [63].

[1] See the debate competition guidelines of the English Speaking Union http://www.esu.org/

[2] http://debategraph.org/

[3] https://www.truthmapping.com/

[4] Crismore et al. (1993) define metadiscourse as "linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organize, interpret and evaluate the information given", e.g. Shifting Topic, Marking Asides, etc.

Table 1: Properties of persuasive public speech (as judged by humans) and their lexico-syntactic, acoustic-prosodic and motion correlates as observed in previous empirical studies.

Delivery aspects	Performance strategy	Correlates		
		linguistic	acoustic-prosodic	visible body movements
Audibility	Adequate voice volume Appropriate argumentation pace	- number of tokens per second	perceived as normal (60-54 dB) number of syllables per second	- number of beats per second
Engagement	Expressiveness	> repetitions (List of Three) [4] > personal pronouns density[49] < information density and redundancy [38, 61]	variations in pitch range [60] > standard deviation in pitch [49, 60]	open gestures (palm) appropriate gesticulation
Conviction	Clear articulation and fluency Adequate prominence and focus, topic-comment structuring	no disfluences and hesitations [7, 63] no false start [49] topicalization, passivization, it- and wh-cleft discourse structuring or meta-discoursive acts [38, 61]	fraction of voiced/unvoiced frames frequent voice breaks > pitch range; > mean pitch; > intensity [7, 21, 41, 49, 61] emphatic accents [39]	hand & arm position posture (e.g. no slouching) adequate beat gestures iconic & metaphoric gestures
Authority	Adequate grouping & phrasing	clear syntactic structures,phrasing, chunking [61]	slowing down speech rate [55, 61] pausing [56, 61, 73]	confident posture
Likability	Express respect and friendliness	sentiment vocabulary, e.g. affect dictionary [72] sentiment shifters; offensive language use [70]	pitch register	eye contact, smiling

Effects of audio-visual prosody on the perception of information status related to focus and prominence have been also studied. For example, investigating *visual beats* it has been concluded that if observers see a visual beat they perceive a corresponding phrase as more prominent [28]. We may expect that prosodically prominent phrases when accompanied by gestures will intensify the assertiveness and persuasion effect of the debate arguments. Good debaters that score high on expression and delivery demonstrate a clear awareness of rhetoric and attempt to engage an audience. They make use of direct eye contact, body language and emotive language.[5] Persuasive debate performance may be linked to dominance. Crossing the arms, stemming the hands on the hip or touching one's neck most effectively influence dominance perception [58].

In summary, 5 global aspects are to be considered: Audibility, Engagement, Conviction, Authority and Likability (AECAL). Although it is often difficult to define clear properties of good debate or public speaking, there are certain linguistic, prosodic and body language features that correlate with human judgments of such behaviour. Debaters make use of these features, employing Frequency, Effort and Production Codes [20], and respecting general Conversational Maxims [19] as well as maxims related to intonational meaning [21], which enables explaining perceptive regularities and to formulate argumentation strategies that trainees may follow to deliver convincing debate performance, and that can be used for its assessment. Table 1 summarizes previous findings on correlations observed between linguistic, acoustic, prosodic speech and visible body movement properties, and human judgments of a 'good rhetoric' linked to AECAL aspects.[6] The presented correlates are not only powerful communicative tools used by skilled debaters to persuade their audience, but they also influence discourse processing to a great extent (see e.g. [14, 47, 71]).

3 DATA COLLECTION AND ANALYSIS

3.1 Scenario

An important step in the design of interactive human-computer systems is to model natural human dialogue behaviour based on the analysis of examples of such behaviour. The specific setting considered for the data collection involves a debate scenario about anti-smoking legislation in Greece. The initial proposal for a smoking ban is supported by the proposing (governmental) party. The goal for the proposer is when aiming at a majority vote to agree on as few amendments as possible.

Our core data collection activity involved debate *trainees*, school children aged 14-15 years who have been exposed to little debate training. Prior to the session each participant was given a set of minimal goals concerning: (1) the total ban on smoking in public spaces, (2) limiting youth access to tobacco products, (3) improving the effectiveness of anti-smoking campaign and (4) raising taxes on tobacco products. Participants were not allowed to disclose their goals to the other parties prior to the interaction. Three human tutors evaluated debate performance.

The collected data consists of 12 sessions with a duration of 2.5 hours, comprising 400 arguments (Argumentative Discourse Units, ADUs[7]) from 6 different bilingual English/Greek speakers, referred to further as the Debate Trainees Corpus (DTC). Each video-recorded debate session involved a pair of participants: one assigned the role of proposer, the other the role of either moderate or conservative opponent. Participants' movements were tracked with two Kinect cameras; speech signals were recorded using a Tascam portable digital recorder. Audio, video and Kinect streams were synchronized based on absolute time stamps with frames of equal 33ms size. Participants' speech was transcribed by correcting the Kaldi-based Automatic Speech Recognizer [45] output.

Trainee's debate performance was compared with those of *skilled* debaters, members of the Youth Parliament, who enjoyed extensive training in a debate school (e.g. the Debate Academy of the English Speaking Union[8]), and *professional* world-class debaters who have made a successful political career. *UK Youth Parliament* (UKYP)[9] debates and the collection of the *American Presidency Project* (APP)[10] were used as benchmarks[11]. The UKYP data comprises three debate sessions with a total duration of 3 hours, and consists of 118 arguments from 35 different speakers, aged 11-18, addressing various youth related current affairs topics. From APP, we selected two

[5] See http://www.esu.org/_data/assets/pdf_file/0011/16202/ESU_Debate_Challenge_2017_v2.pdf

[6] The matrix presented in Table 1 is rather simplified. In reality the mapping is not 1:1 and cross-factor dependencies exist.

[7] For more details on segmentation and annotation performed, we refer to [43].

[8] http://www.esu.org/our-work/debate-academy

[9] http://www.ukyouthparliament.org.uk/

[10] http://www.presidency.ucsb.edu/index.php

[11] It should be noticed that for these corpora maby prosodic features were not considered in the analysis due to the low quality of audio recordings. Kinect tracking data is also not available.

Table 2: Distribution of relation tags in the data, in % (in brackets DTC figures; * = relations marked in the CE-EMNLP-2015 training data.

Top level	**Support**		***Non-Support***
Coarse level	*Contingency*	*Evidence*	
Fine-grained relations			
justification	34 (32)	-	-
reason	4.5 (2.5)	-	-
motivation	12 (12.5)	-	-
exemplification	-	7.5 (11)	-
explanation	-	7 (9)	-
exception	-	1 (1)	-
no-relation	-	-	34 (32)
Total:	50.5 (47)	15.5 (21)	34 (32)
*study**	-	18	-
*study/expert**	-	5	-
*expert**	41	-	-
no-relation	-	-	36
Total:	41	23	36

presidential and one vice-presidential debate sessions on multiple current affairs topics with a total duration of 4.5 hours. UKYP debates are mostly prepared speeches, while the DTC and APP debates are largely impromptu speeches.

We also used two text corpora: (1) the in-domain Forum Corpus constructed from smoking ban discussions on online debate forum[12] containing 84 arguments; and (2) the CE-EMNLP-2015 corpus [48] comprising 2294 Wikipedia arguments.

3.2 Annotation

The recorded and collected data was segmented and annotated with dialogue act information following the ISO 24617-2 standard [24]. Segments were assigned one or more communicative function in nine ISO dimensions and linked according to ISO 24617-8 by discourse relations [10]. The ISO 24617-2 taxonomy [24] distinguishes 9 dimensions, addressing information about a certain *Task*; the processing of utterances by the speaker (*Auto-feedback*) or by the addressee (*Allo-feedback*); the management of difficulties in the speaker's contributions (*Own-Communication Management*) or that of the addressee (*Partner Communication Management*); the speaker's need for time to continue the dialogue (*Time Management*); the allocation of the speaker role (*Turn Management*); the structuring of the dialogue (*Dialogue Structuring*); and the management of social obligations (*Social Obligations Management*). A good inter-annotator agreement has been reached between two trained annotators ranging from 0.85 to 0.95 in terms of Cohen's kappa for the segmentation task, 0.59 to 0.81 for the assignment of discourse relation links and 0.71 to 0.86 for the classification of discourse relation types. In the data, more than 41.4% of the dialogue acts performed by the debaters are Inform acts, which are often connected by discourse relations forming an argument (ADU). Small portions of Set Questions (3.4%) and Agreements or Disagreements (1.7%) are observed. Other dialogue acts are concerned with Turn Management (22.7%); Time Management (21.1%); Own Communication Management (7.3%); Social Obligation Management (1.2%); and Discourse Structuring acts (10%). Units with argumentation relevant discourse relations such as *Contingency* and *Evidence* were extracted and spanned at the top level into a *Support* class.

[12] http://www.debate.org/

Table 3: Distribution of annotated gesture events in terms of their relative frequency (in%) and proportion of frames (in %).

Type of gesture		Relative frequency (in %)	Proportion of total 7074 frames (in %)
Beats	all categories	59.55	27.04
	prominence intensifier	69.76	68.90
	new topic/theme marker	3.26	3.45
	meta-discursive act marker	17.67	16.36
	phrase/boundary marker	9.31	11.29
Adaptors		14.96	18.80
Iconic		2.22	1.37
Deictic		2.22	1.84
Emblem		0.55	0.24
Unclassified gesture event		20.50	50.7

At the fine-grained level, Contingency relations were subdivided into *justification, motivation* and *reason* relations; Evidence ones into *exemplification, explanation* and *exception*. Non-argumentative statements were labeled as having *Non-Support* relations. Table 2 shows the distribution of relation tags in the collected data.

3.3 Multimodal Features

Features extracted and computed from the data are related to debaters' linguistic, acoustic and non-verbal behaviour.

As for *linguistic* features, we computed bag-of-words (bow) vectors, bi- and 2-skip-bigrams, word pairs, doc2vec [35] and occurrences of modal verbs in a feature vector. As the basic and most easily computable feature, bow is used to compute baselines. As for doc2vec, a 100 dimensional vector was computed. The length of an argument (#tokens) and number of syntactic and semantic constituents (#chk) served as *structural features*.[13]

For each frame prosodic properties were computed automatically using PRAAT [5] such as minimum, maximum, mean, and standard deviation in *pitch*, *energy*, *voicing* and speaking rate.[14]

Visible movement features were extracted or computed from the Kinect output and comprise overall gesture, gesture stroke and retraction phase *duration* for each hand; *handedness* for right, left or both hands movements; *X, Y, Z coordinate values* for each hand and frame; and *X, Y, Z coordinate values* for gesture stroke and retraction phase for each hand. The prosodic and visual movements history of 5 previous frames was encoded in a feature vector. The participant's co-speech gestures were manually annotated considering beats, iconics, deictics, emblems and adaptors [26, 34].[15] The distribution of annotated gesture events is shown in Table 3 in relative frequencies and in proportional gesture events duration. Beat gestures are the most frequent gesture type, mostly used as prominence markers.

4 ARGUMENT STRUCTURE

When involved in argumentative discussions, debaters plan the structure of their arguments. For this, they typically apply one of the known techniques discussed in Section 2. Knowing these

[13] The SENNA parser [3] was used to identify semantic roles and modifiers of time, location, manner, attributes and negations.

[14] We computed both raw and normalized versions of these features. Speaker-normalized features were obtained by computing z-scores (z = (X-mean)/standard deviation) for the feature, where mean and standard deviation were calculated from all functional segments produced by the same speaker in the debate session. We also used normalizations by the first speaker turn and by prior speaker turn.

[15] Two independent annotators reached a good inter-annotator agreement on average in terms of Cohen's kappa of 0.64 [12].

Table 4: Classification results on best feature combinations for three classification granularity levels using independent and cascade classification procedures and different corpus combinations, in terms of accuracy (in %). * = differs significantly from the baseline according to two-sided t-test, t < .05

Classification	Feature	DTC		DTC + Forum		DTC + Forum + CE-EMNLP-2015	
		independent	cascade	independent	cascade	independent	cascade
TOP	bow (baseline)	0.58	-	0.60	-	0.49	-
	skipgrams + #chks	0.77*	-	0.79*	-	0.59*	-
COARSE	bow (baseline)	0.44	0.46	0.45	0.46	0.40	0.40
	skipgrams + #chks or #args	0.53*	0.54*	0.47*	0.49*	0.52*	0.52*
FINE	bow (baseline)	0.45	0.45	0.39	0.40	0.40	0.40
	bigrams +skipgrams + #chks	0.45	0.46*	0.30	0.30	0.40	0.40

Table 5: Correlations between computed argument features and average Likert scores assigned by human judges in 'reading an argument' and 'hearing an argument' experiments. * = differs significantly from zero according to two-sided t-test, t < .05

Feature type	Pearson correlation coefficient (R)	
	'reading an argument'	'hearing an argument'
syntactic constituents	0.62	0.02
semantic constituents	-0.08	-0.42*
number of tokens	-0.09	-0.38*
number of referring expression (claim)	-0.62*	-0.17
number of referring expression (evidence)	-0.39*	-0.09
number of referring expression (total)	-0.61*	-0.25
pronoun density	-0.17	0.09
pronoun density (1st person singular & plural)	-0.17	0.16
lexical density	-0.24	-0.23
disfluency ratio	-	-0.65*

structural patterns helps to automatically identify and classify argument constituents and relations between them. Good debaters are distinguished by concise clear arguments connected by explicitly signaled markers and discourse structuring acts. Exploiting these facts, many frameworks for argumentative discourse analysis aimed to capture discourse coherence by integrating discourse segments into larger structural units. One of the first approaches, Teufel's Argumentative Zoning [59], is based on computing linguistic features like sentence length, word forms or cue phrases for Mann's and Thompson's (1988) rhetorical relations in scientific texts (F-scores of 0.46 obtained). More recently, Mochales-Palau and Moens (2009) divide the problem into sub-problems: (1) the identification of arguments, (2) the classification of internal argument structure, and (3) the recognition of relations between the different arguments. They use Support Vector Machines (SVM, Boser et al., 1992) trained on features such as sentence length, main verb type and tense, rhetorical patterns, clause-internal citations, etc. They obtained F-scores of 74.07% for conclusion and 68.12% for premise classification in the European Court of Human Rights corpus.

We conducted series of stratified 5-fold cross-validation, SVM-based learning experiments to (1) classify relations in isolation and in groups at various granularity levels; (2) assess the features importance; (3) combine different corpora in training sets, and (4) apply various classification procedures.

The results are summarised in Table 4. As for features, all but *doc2vec* and the presence of a *modal verb* outperform the baseline for the *top* level classification, discriminating well between the *Support* and *Non-Support* classes. The classifiers performance on *doc2vec* features was rather poor when trained on the DTC data. It can be explained by the fact that the data gets too sparse to create a vector space representative for the target domain. When trained and evaluated on the CE-EMNLP-2015 data, classifiers using the doc2vec feature achieved an accuracy of 0.8 for top level classification. Argument wording is important for the argument structure mining task. The best performance has been achieved when combining *skip-grams* and *chunking* information. This suggests that argument claims, premises and conclusions may have distinctive syntactic structures.

A performance increase is observed for both in-domain corpora. The classification, by contrast, does not benefit from in- and out-of-the domain data combination. Testing CE-EMNLP-2015 prediction models on the target DTC data showed performance below the baseline. Thus, the data quality matters, not only its quantity.

Three sets of classifiers were trained to classify discourse relations at the top and coarse level, and fine-grained classes independently. Higher level class predictions were added as features to classify instances at the lower more fine-grained level (cascading). Cascade classification outperforms the independent one.

5 ARGUMENT QUALITY

The assessment of argument quality is often based on formally defined argumentation schemes. Existing systems, e.g. the Carneades system[16], are good in reasoning tasks within abstract argumentation frameworks, but require externally specified task models to be successful in checking the logical consistency of arguments within a specific domain. Moreover, argument quality may be a rather subjective metric and is primarily concerned with clarity, coherence and comprehensibility. To assess argument quality we propose a method that relates data features to human judgments. To establish a set of operational criteria for automatic debate argument quality assessment, we conducted two studies measuring correlations between basic features extracted or easily computed from data and human judgments. For the first study 16 naive subjects were asked to judge the quality of 29 randomly selected transcribed arguments of debate trainees ('reading an argument') on a 5-point Likert scale [31]. In the second experiment, 15 different naive subjects listened to 23 recorded arguments selected from the previous set ('hearing an argument'). The subjects were asked to rate the quality of the arguments on the same scale. A moderate inter-rater agreement was reached in terms of Krippendorff's alpha [29], reaching 0.56 for the first and 0.59 for the second experiment. We used a subset of classification features such as the #tokens, and semantic and syntactic #chk to measure syntactic and semantic argument complexity. Additionally, the number of referring expressions (personal, temporal, locative and discourse deictics), pronouns density, disfluency ratio and lexical density were computed. The latter is a

[16] http://carneades.github.io/

Table 6: Summary of observations on argument production fluency for confidence and clarity assessment of trainees vs skilled vs professional debaters in prepared vs impromptu speech.

Linguistic/prosodic/ temporal phenomenon	Trainees *impromptu speech*	Skilled debaters *prepared speech*	Professional debaters *impromptu speech*
Ratio filled pauses / total ADU tokens	from 0.10 to 0.19	0.0	from 0.01 to 0.02
Ratio duration filled pauses /total ADU duration	from 0.3 to 0.4	0.0	close to 0.0
Ratio restarts /total ADU tokens	from 0.05 to 0.1	0.0	close to 0.0
Ratio retractions/total ADU tokens	from 0.08 to 0.24	0.0	0.0
Speaking rate in syllables/sec	from 1.2 to 10.5	from 0.9 to 5.7	from 2.0 to 4.2
Ratio silent pauses/ ADU clauses	from 0.9 to 1.7	from 1.7 to 1.9	from 2.1 to 2.95

simple metric measuring the complexity of information provided by an argument expressed as the ratio of function to content words. Subsequently, Pearson's correlation tests were performed between the mean Likert scores and the numerical feature values [40].

As the results suggest, the presence of referring expressions has significant negative effects, see Table 5. This means that the more referring expressions an argument contains the harder it was for the subjects to understand it, especially if those expressions occur in the claim. These effects were not observed for 'hearing an argument'. This may be explained by the fact that when reading an argument, the reader selects his own focus of attention, which is potentially different from the one of the debater. Another between-conditions difference is observed in the perceived argument complexity in terms of number of semantic constituents and argument length. The measured effects suggest that long and complex arguments are more difficult to comprehend when listening to them than when reading them. This was to be expected, as numerous studies on comprehension of spoken and written sentences across multiple languages and domains have shown similar effects, see e.g. [8].

The fact that referring expressions are perceived as complications for argument understanding most probably means that the argument quality is context dependent. For instance, we observed that if co-referential expressions occur in an evidence utterance, an argument was still successfully interpreted as good support for a claim. Failing to identify antecedents from the previous context for the referents in the argument claim, by contrast, seriously complicates the overall argument comprehension. As a metric for the clarity of an argument and speaker's confidence, the disfluency ratio showed a moderate to strong negative effect.

6 ARGUMENT DELIVERY

To assess argument delivery aspects we performed a series of experiments of different types, including observational studies from the collected data and correlation experiments that measure the perceived strength of confidence and intensification effects.

Observational studies involved straightforward measures describing the basic data features and comparing them to those of benchmarks. To detect regularities, we calculated feature distributions (i.e. relative frequencies) and ratios. Skilled professional debaters were observed to avoid filled pauses, frequent editing expressions and repairs. In impromptu professional speeches such phenomena were rather rare, while in prepared speech they were absent. Disfluencies were of short duration. Generally, professionals seem to prefer silent pauses to filled ones. Well-timed pauses are used for prominence and at transition places to a new segment/topic, making the speaker perceived as more confident and assertive. Skilled professional debaters use measured speaking rate,

Table 7: Correlations between computed mutimodal argument features and confidence level scores assigned by three human debate coaches. r = Pearson correlation coefficient; * = differs significantly from zero according to two-sided t-test, t < .05

Audio-visual features	r
mean pitch	0.07
standard deviation pitch	-0.269*
max pitch	0.318*
fraction of unvoiced frames (FoUF)	-0.258*
number of voice breaks (NoVB)	-0.356*
mean intensity	0.262*
gaze aversion	-0.42*
beat gesture (> 20/*min*)	- 0.61*
invisible hands (incl. adaptors)	-0.59*
random posture shifts (> 40/*min*)	- 0.87*

whereas the performance of trainees is less balanced in this respect. Table 6 summarizes our findings for linguistic, prosodic and temporal aspects of fluent confident speech. We report the observed lowest and upper values, and do not average over speakers.

Correlation experiments based on bivariate Pearson tests measured the significance of linear relationship between the trainees debate performance and the judgments of three professional debate coaches who assigned a persuasiveness level ranging from 0 (very not-confident performance) to 5 (very confident performance). Standard deviation in pitch can be observed to have a strong negative correlation with perceived speaker confidence: higher standard deviation is perceived as lower confidence (see Table 7). This is not entirely in line with the conclusions in [49], where a higher standard deviation in pitch is explained as a signal of expressiveness and positively correlating with charisma judgments, although the relation with human perception of charismatic speech may differ from that of confident speech. Significant positive effects of maximum pitch and mean intensity are found and explained by the fact that confident speakers do stress important and contrastive information and speak 'up'. The debater is perceived as less confident when he uses a higher number of voice breaks and a significant portion of unvoiced frames is detected. As for visible movements, significant strong negative effects of extensive gaze aversion, frequent abrupt beat gestures, 'invisible hands' (e.g. in pockets or behind the back) and frequent random posture shifts ('dancing') were found. In sum, clear, fluent speech of balanced speed and with meaningful argument-internal pauses, open body position without excessive gesticulation and shifts is perceived as confident and persuasive.

7 VIRTUAL DEBATE COACH: DESIGN AND EVALUATION

The Virtual Debate Coach (VDC) was designed with the functionality described in the data collection. The VDC "hears" and "sees" a

Table 8: Overview of the feedback on inappropriate debate behaviour as detected by the VDC system.

Coaching aspect	Criteria	Feature
Argument structure	missing/unmarked reason/evidence relations	no discourse marker detected
	abrupt and frequent interruptions	overlapping speech of > 500*ms*
Argument quality	lengthy arguments	turn that > 1 minute
	irrelevant arguments	> 50% out-of-vocabulary tokens, computed from the domain language model
	high number of syntactic/semantic chunks	> 24 syntactic constituents per ADU
	high number of referring expressions	> 7 referring expressions per ADU
Argument delivery	speech fluency: number of ADU-internal pauses	> 7 silent pauses that are > 200*ms* per ADU
	speech volume: not adequate speech volume	too loud (> 60 dB); too soft (< 30 dB)
	hands/arms position: arms crossed, hands in the pockets, behind the back	> 60 missing Kinect frames for hand joints per ADU
	gesticulation: number of hand movements	> 70% of unclassified gesture events

wide range of signals, interprets them, and identifies relevant markers as described above and simmarized in Table 8. The trainee is expected to deliver better performance and gain confidence through practicing debates, and through the VDC visual or spoken feedback. Figure 1 shows the VDC architecture and processing workflows.

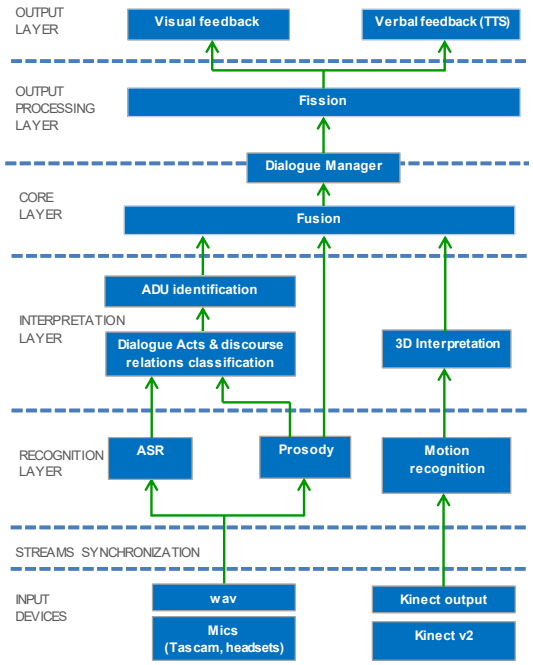

Figure 1: Virtual Debate Coach architecture. From bottom to top, signals are received through input devices, and processed by tailored modules. After interpretation concerned with dialogue acts, relation classification, and ADU identification, semantic representations from different modalities are fused and passed to the Dialogue Manager for context model update and feedback generation. The generated feedback is rendered in different output modalities.

Speech signals are recorded from multiple sources, such as wearable microphones, headsets for each dialogue participant, and an all-around microphone placed between participants. The speech signals are passed to Automatic Speech Recognition (ASR). The Kaldi-based ASR component incorporates acoustic and language models developed using various available data sources. In total, about 759 hours of data has been used to train an acoustic model[17]. The collected Metalogue DTC and in-domain Forum data is used for language model adaptation. The ASR performance is measured at 34.4% Word Error Rate (WER), see [54]. The ASR outputs the 1st best word sequence. Prosodic properties related to voice quality, fluency, stress and intonation were computed as described above.

The ASR output is used to classify dialogue acts and discourse relations between them. For the discourse-based argument structure identification, the performance of 0.54 in terms of F-scores was achieved. For the recognition of the intentions encoded in debater' utterances, Support Vector Machine [6], Logistic Regression [75], AdaBoost [76] and the Linear Support Vector Classifier (LinearSVC) [67] were applied with F-scores range between 0.83 and 0.86 [1].

Kinect tracked data is used to detect hand/arm co-speech gestures[18] and their types (see Table 3 for the distribution of gesture events). SVM and Gradient Boosting [18] classifiers were trained and achieved F-scores of 0.72 [46]. The motion interpretation component related to hand/arms position detection of the designed Presentation Trainer ([52, 66]) is integrated into the VDC system.

The system includes a Fusion component, which combines the modality-specific analyses into a fused representation of debater's actions related to argument structure, quality and delivery aspects. For instance, prosodic and motion tracking information has been combined to interpret the status of information conveyed in an argument. Exploiting the fact that pitch-accented tokens often coincide with focus, topic and contrast, and if accompanied by a beat gesture are perceived as even more prominent, we identified 95% of all beat gesture events produced around intensity peaks. The fusion module also incorporates an SVM-based classifier that operates on prosodic and motion features, and predicts the persuasiveness level of an argument with an accuracy of 71% [44].

Given the system's understanding of the trainee's behaviour, the VDC task is to perform tutoring interventions by informing the trainee of a mistake or proposing corrections (or to provide positive feedback). The performance on this task requires immediate real-time feedback, often called 'in-action' feedback (Schön, 1983) on the three aspects mentioned above. The Dialogue Manager (DM), designed as a set of processes (threads), receives data, updates the information state and generates the VDC feedback, see [32].

Given the feedback dialogue acts provided by the Dialogue Manager, Fission module generates system responses, splitting content

[17]Examples of resources are: the Wall Street Journal WSJ0 corpus https://catalog.ldc.upenn.edu/ldc93s6a, HUB4 News Broadcast data https://catalog.ldc.upenn.edu/ldc98s71, the VoxForge corpus http://www.voxforge.org/, etc.

[18]Co-speech gestures are visible hand/arm movements produced alongside speech and are interpretable only through their semantic relation to the synchronous speech content.

Table 9: Results evaluating effectiveness, efficiency and user satisfaction. M = Mean; SD = Standard Deviation.

Usability metric	Example of trainee's survey entry	M	SD
effectiveness (task success)	I completed my task successfully	4.95	0.72
effectiveness (task quality)	I achieved all my goals	3.35	0.92
efficiency	The system feedback was mostly timely	3.4	1.05
	System feedback was valuable	3.7	0.91
	System feedback made me more aware of my performance	3.45	1.2
	System provided enough feedback	3.07	1.4
satisfaction, (QUIS) [11]	I found the interaction with the system natural	3.95	1.15
	I found the interaction with the system engaging	4.7	0.75
	I found the interaction with the system useful	3.95	0.84
	I would use the system in my training routine	4.37	0.86

into different modalities, such as Avatar, voice (TTS) and visual feedback for tutoring interventions. At the end of each debate session, summative feedback is generated summarizing the number of arguments, hesitations, interruptions, editing expressions, etc. It should be noted here that all messages exchanged between modules are in the standard TEI [23] and ISO DiAML [9] formats.

We performed trainee-based evaluation experiments involving 40 trainees (male female aged between 14 and 20 years). We did not aim at the trainees learning gain assessment which has been performed in a separate study involving the complete 'learner journey' scenario, see also [64] and [27]. The main evaluation goal of the presented study was to assess system performance in the trainee-based setting, including assessing types, granularity, amount and timing of coaching interventions expected to lead to the best learning outcome. For this, participants debated in pairs as described in Section 3. A debriefing stage included filling in questionnaires and discussion rounds with trainees and tutors. Questionnaires were constructed in such a way that, along with overall trainee satisfaction, we could also link their judgments to the system's coaching interventions. Trainee judgments were presented in a 1-5 Likert scale. Each session lasted 60-90 minutes including preparation, interaction and filling in a questionnaire. The discussion round involved all participants and tutors after all sessions are completed.

The VDC generated real-time 'in-action' feedback on presentational and interactive aspects such as speech volume, speaking rate, hand and arm position, posture shifts, and turn taking and time management behaviour, i.e. interruptions, overlapping speech and arguments longer than >1 minute were discouraged. Full session recordings, system recognition and processing results, as well as the generated 'in-action' feedback were logged and converted to `.anvil` format for using Anvil tool to view, browse, search, replay and edit debate sessions. This allows automatic generation of the VDC summative feedback to be discussed in 'about-action' feedback sessions. Moreover, the implemented prediction models can be edited by debaters and tutors on the fly, and corrected annotations can be used to retrain the system.

Table 9 summarizes results and shows consistently positive participant feedback for almost all the questions, however, with different deviations from the mean. High task completion rate along with positive effect on skill training is reported. Trainees indicated however that system feedback was sometimes hard to interpret. Most participants found that the system generated too much feedback; such a large amount was difficult to process and distracted from the debate interaction. Trainees also expected more real-time feedback and summative feedback on learning progress.

8 CONCLUSIONS AND FUTURE WORK

This paper presents an approach to the assessment of natural multimodal argumentative behaviour based on the defined set of criteria used to explain observed regularities and to define rules, strategies and constraints for the generation, assessment and correction of trainees' debate performance. Various experiments supported fairly reliable identification of multimodal markers, and linked them to assessments of argument structure, quality and its delivery aspects.

We observed that linguistic features (i.e.n-gram of various size and types in combination with syntactic information), multimodal in-domain corpora and classification procedures resulted in the best performance on an argument structure mining task. Results of the argument quality experiments showed that argument comprehensibility is affected by the number of referring expressions, information complexity, and presentation fluency. Presence of intensification and segmentation markers, position and movements of hands/ams and certain postures may affect the perception of the clarity, persuasiveness, and credibility of debaters.

The Virtual Debate Coach that we designed and implemented on the basis of these theoretical frameworks and empirical findings was positively evaluated in a trainee-based setting.

The ambitious vision of the VDC presents a significant number of challenges. A fully automatic system that is able to understand natural arguments in a debate accurately enough to achieve human-like performance has not been yet achieved due to certain limitations in sensor tracking, speech recognition, and natural language processing technologies. Also since a data-oriented approach for modelling of many debate phenomena has been deployed, the currently available quantity and quality of multimodal data are insufficient for training statistical machine learning algorithms.

There is a lot of room for further research. Our main goal is to advance in achieving *immersive* coaching, when the system will enter, exit and re-enter different modes, e.g. monitoring, mirroring, exercising, reflecting, guiding and freestyle modes. Integrated immersive feedback will be enabled by multiple interactive modalities and media including visual, auditory, typed and handwritten presentation. We also are planning to advance argument assessment by considering more elaborate argument contexts and further qualitative linguistic and multimodal features, e.g. related to information density and complexity, accounting for surprisal, and by incorporating additional sensing devices.

ACKNOWLEDGMENTS

This research was partly funded by the EU FP7 Metalogue project, under grant agreement number: 611073. We are also very thankful to anonymous reviewers for their valuable comments.

REFERENCES

[1] D. Amanova, V. Petukhova, and D. Klakow. 2016. Creating Annotated Dialogue Resources: Cross-Domain Dialogue Act Classification. In *Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Paris, Portorož, Slovenia, 111–117.

[2] K. Ashley, N. Pinkwart, C. Lynch, and V. Aleven. 2007. Learning by Diagramming Supreme Court Oral Arguments. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL '07)*. ACM, Stanford, California, 271–275.

[3] M. Bansal, K. Gimpel, and K. Livescu. 2014. Tailoring Continuous Word Representations for Dependency Parsing.. In *ACL (2)*. Association for Computational Linguistics, Baltimore, US, 809–815.

[4] A. Beard. 2002. *The language of politics.* Routledge, London.

[5] P. Boersma and D. Weenink. 2009. Praat: doing phonetics by computer. Computer program. (2009). Available at http://www.praat.org/.

[6] B. Boser, I. Guyon, and V. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual workshop on Computational learning theory*. ACM, Pittsburgh, PA, USA, 144–152.

[7] D. Braga and M.A. Marques. 2004. The pragmatics of prosodic features in the political debate. In *Speech Prosody 2004, International Conference*. ISCA Special Interest Group on Speech Prosody, Nara, Japan, 321–324.

[8] A. Buchweitz, R. Mason, L. Tomitch, and M. Just. 2009. Brain activation for reading and listening comprehension: An fMRI study of modality effects and individual differences in language comprehension. *Psychology & neuroscience* 2, 2 (2009), 111.

[9] H. Bunt, J. Alexandersson, J.-W. Choe, A. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *LREC*. ELRA, Paris, Istanbul, Turkey, 430–437.

[10] H. Bunt and R. Prasad. 2016. ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. ELRA, Paris, Portorož, Slovenia, 45–54.

[11] J. Chin et al. 1988. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, Washington, DC, US, 213–218.

[12] J. Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement* 20 (1960), 37–46.

[13] A. Crismore, R. Markkanen, and M. Steffensen. 1993. Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written communication* 10, 1 (1993), 39–71.

[14] D. Dahan, M. Tanenhaus, and C. Chambers. 2002. Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language* 47, 2 (2002), 292–314.

[15] C. Dede. 2009. Immersive interfaces for engagement and learning. *Science* 323, 5910 (2009), 66–69.

[16] C. D'Souza. 2013. Debating: a catalyst to enhance learning skills and competencies. *Education+ Training* 55, 6 (2013), 538–549.

[17] J.B. Freeman. 2011. Argument structure: representation and theory. In *Argumentation Library*. Vol. 18. Springer, Berlin.

[18] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (2002), 367–378.

[19] P. Grice. 1975. Logic and conversation. In *Perspectives in the Philosophy of Language: A Concise Anthology*, Robert J. Stainton (Ed.). Broadview Press, Ontario, Canada, 41–58.

[20] C. Gussenhoven. 2002. Intonation and interpretation: Phonetics and phonology. In *In Proceedings of the Speech Prosody 2002 International Conference*. Laboratoire Parole et Langage, SProSIG, Aix-en-Provence, France, 47–57.

[21] J. Hirschberg. 2002. The pragmatics of intonational meaning. In *Speech Prosody 2002, International Conference*. Laboratoire Parole et Langage, Aix-en-Provence, France, 65–68.

[22] T. Hughes, J. Flatt, B. Fu, C. Chang, and M. Ganguli. 2013. Engagement in social activities and progression from mild to severe cognitive impairment: the MYHAT study. *International psychogeriatrics* 25, 04 (2013), 587–595.

[23] ISO. 2006. *TEI-ISO 24610-1:2006 Language resource management: Feature structures, Part 1: Feature structure representation.* ISO, Geneve.

[24] ISO. 2012. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2.* ISO Central Secretariat, Geneva.

[25] S.L. Johnson. 2009. *Winning Debates: A Guide to Debating in the Style of the World Universities Debating Championships.* International Debate Education Association, Brussels, Belgium.

[26] A. Kendon. 2004. *Gesture: visible action as utterance.* Cambridge University Press, Cambridge.

[27] D. Koryzis, V. Svolopoulos, and D. Spiliotopoulos. 2016. Metalogue: A Multimodal Learning Journey. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, Corfu, Island, Greece, 48.

[28] E. Krahmer and M. Swerts. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language* 57, 3 (2007), 396–414.

[29] K. Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality and Quantity* 38:6 (2004), 787–800.

[30] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff. 2001. A cross-media presence questionnaire: The ITC-Sense of Presence Inventory. *Presence* 10, 3 (2001), 282–297.

[31] R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 22, 140 (1932), 1–55.

[32] A. Malchanau, V. Petukhova, H. Bunt, and D. Klakow. 2015. Multidimensional dialogue management for tutoring systems. In *Proceedings of the 7th Language and Technology Conference (LTC 2015)*. Faculty of Mathematics and Computer Science of the Adam Mickiewicz University, Poznan, Poland, 482–486.

[33] W. Mann and S. Thompson. 1988. *Rhetorical structure theory: toward a functional theory of text organisation.* MIT Press, Cambridge, MA.

[34] D. McNeill. 1992. *Hand and mind: What gestures reveal about thought.* University of Chicago Press, Chicago, Illinois.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* 1301.3781 (2013).

[36] R. Mochales-Palau and M-F. Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009)*. ACM, Barcelona, Spain, 98–109.

[37] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed. 2007. Automatic Detection of Arguments in Legal Texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL '07)*. ACM, Stanford, California, 225–230.

[38] R. Nir. 1988. Electoral rhetoric in Israel - the television debates. A study in political discourse. *Language Learning* 38:2 (1988), 187=208.

[39] E. Novák-Tót, O. Niebuhr, and A. Chen. 2017. A gender bias in the acoustic-melodic features of charismatic speech?. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. International Speech Communication Association (ISCA), Baixas, France, Stockholm, Sweden, 2248–2252.

[40] K. Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58 (347-352) (1895), 240–242.

[41] A. Pejčić. 2014. Intonational Characteristics of persuasiveness in Serbian and English Political debates. *Nouveaux Cahiers de Linguistique Française* 31 (2014), 141–151.

[42] A. Peldszus and M. Stede. 2013. From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7(1) (2013), 1–31.

[43] V. Petukhova, A. Malchanau, and H. Bunt. 2016. Modelling argumentative behaviour in parliamentary debates: data collection, analysis and test case. In *Principles and Practice of Multi-Agent Systems. Lecture Notes in Artificial Intelligence*, M. Baldoni, C. Baroglio, F. Bex, F. Grasso, N. Green, M. Namazi-Rad, M.-R.and Numao, and M.T. Suarez (Eds.). Springer, Berlin, 26–46.

[44] V. Petukhova, M. Raju, and H. Bunt. 2017. Multimodal markers of persuasive speech : designing a Virtual Debate Coach. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. International Speech Communication Association (ISCA), Baixas, France, Stockholm, Sweden, 142–146.

[45] D. Povey. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Big Island, HI, US.

[46] M. Raju. 2016. *Automatic Detection and Classification of Beat Gestures in Argumentative Discourse.* Master's thesis. Saarland University, Germany.

[47] S. Repp and H. Drenhaus. 2015. Intonation influences processing and recall of left-dislocation sentences by indicating topic vs. focus status of dislocated referent. *Language, Cognition and Neuroscience* 30, 3 (2015), 324–346.

[48] R. Rinott, L. Dankin, C. A. Perez, M. Khapra, E. Aharoni, and N. Slonim. 2015. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection.. In *EMNLP*. The Association for Computational Linguistics, Lisbon, Portugal, 440–450.

[49] A. Rosenberg and J. Hirschberg. 2009. Charisma perception from text and speech. *Speech Communication* 51.7 (2009), 640–655.

[50] W. Sadowski and K. Stanney. 2002. Presence in virtual environments. In *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. Hale and K. Stanney (Eds.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 791—806.

[51] S. Sali, N. Wardrip-Fruin, S. Dow, M. Mateas, S. Kurniawan, A. Reed, and R. Liu. 2010. Playing with words: from intuition to evaluation of game dialogue interfaces. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*. ACM, Monterey, CA, US, 179–186.

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

[53] D. A. Schön. 1983. The Reflective Practitioner: How Professionals Think in Action. In *Basic Books*, T. Smith (Ed.). Temple Smith, London.

[54] M. Singh, Y. Oualil, and D. Klakow. 2017. Approximated and domain-adapted LSTM language models for first-pass decoding in speech recognition. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. International Speech Communication Association (ISCA), Baixas, France, Stockholm, Sweden, 2720–2724.

[55] E. Strangert. 1991. Phonetic characteristics of professional news reading. *PERILUS* XII (1991), 39–42.

[56] E. Strangert. 2005. Prosody in public speech: analyses of a news announcement and a Political interview.. In *Proceedings of the 6th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. International Speech Communication Association (ISCA), Baixas, France, Lisbon, Portugal, 3401–3404.

[57] E. Strangert and T. Deschamps. 2006. The prosody of public speech - a description of a project. *Lund Unversity Working Papers* 52 (2006), 121–124.

[58] C. Straßmann, A. von der Pütten, R. Yaghoubzadeh, R. Kaminski, and N. Krämer. 2016. The Effect of an Intelligent Virtual Agent's Nonverbal Behavior with Regard to Dominance and Cooperativity. In *International Conference on Intelligent Virtual Agents*. Springer, Los Angeles, CA, US, 15–28.

[59] S. Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. Dissertation. University of Edinburgh, Edinburgh, Scotland.

[60] P. Touati. 1993. Prosodic aspects of political rhetoric. In *ESCA Workshop on Prosody*. International Speech Communication Association (ISCA), Baixas, France, Lund, Sweden, 168–171.

[61] P. Touati. 2009. Temporal profiles and tonal configurations in French political speech. *Working Papers in Linguistics* 38 (2009), 205–219.

[62] S. Toulmin. 1958. *The Uses of Arguments*. Cambridge University Press, Cambridge, England.

[63] C. Tuppen. 1974. Dimensions of communicator credibility: An oblique solution. *Speech Monographs* 41:3 (1974), 253–260.

[64] J. Van Helvert, V. Petukhova, C. Stevens, H. de Weerd, D. Börner, P. Van Rosmalen, J. Alexandersson, and N. Taatgen. 2016. Observing, Coaching and Reflecting: Metalogue - A Multi-modal Tutoring System with Metacognitive Abilities. *EAI Endorsed Transactions on Future Intelligent Educational Environments* 16, 6 (2016). https://doi.org/10.4108/eai.27-6-2016.151525

[65] J. Van Merriënboer and P. Kirschner. 2012. *Ten steps to complex learning: A systematic approach to four-component instructional design*. Routledge, Oxford, UK.

[66] P. Van Rosmalen, D. Börner, J. Schneider, V. Petukhova, and J. Van Helvert. 2015. Feedback design in multimodal dialogue systems. In *Proceedings of the 7th International Conference on Computer Supported Education*, M. Helfert, M. T. Restivo, S. Zvacek, and J. Uhomoibhi (Eds.). SCITEPRESS, Lisbon, Portugal, 209–217.

[67] V. Vapnik. 2013. *The nature of statistical learning theory*. Springer Science & Business Media, Heidelberg/Berlin, Germany.

[68] A. Vinciarelli, M. Pantic, and H. Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[69] D. N. Walton. 1996. *Argumentation schemes for presumptive reasoning*. Routledge, Oxford, UK.

[70] W. Warner and J. Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, Montreal, Canada, 19–26.

[71] D. Watson, M. Tanenhaus, and C. Gunlogson. 2008. Interpreting pitch accents in online comprehension: H* vs. L+ H. *Cognitive Science* 32, 7 (2008), 1232–1244.

[72] C. Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports* 105, 2 (2009), 509–521.

[73] A. Wichmann. 2002. Attitudinal intonation and the inferential process. In *Speech Prosody 2002, International Conference*. Laboratoire Parole et Langage, Aix-en-Provence, France, 11–16.

[74] B. Woods, E. Aguirre, A. Spector, and M. Orrell. 2012. Cognitive stimulation to improve cognitive functioning in people with dementia. *Cochrane Database Syst Rev* 2, 2 (2012).

[75] H.-F. Yu, F.-L. Huang, and C.-J. Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning* 85, 1-2 (2011), 41–75.

[76] J. Zhu, H. Zou, S. Rosset, T. Hastie, et al. 2009. Multi-class adaboost. *Statistics and its Interface* 2, 3 (2009), 349–360.