INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multimodal markers of persuasive speech : designing a Virtual Debate Coach

Volha Petukhova[1], *Manoj Raju*[1], *Harry Bunt*[2]

[1]Spoken Language Systems Group, Saarland University, Germany
[2]Tilburg Center for Communication and Cognition, Tilburg University, The Netherlands
{manoj.raju;v.petukhova}lsv.uni-saarland.de, harry.bunt@uvt.nl

Abstract

The study presented in this paper is carried out to support debate performance assessment in the context of debate skills training. The perception of good performance as a debater is influenced by how believable and convincing the debater's argumentation is. We identified a number of features that are useful for explaining perceived properties of persuasive speech and for defining rules and strategies to produce and assess debate performance. We collected and analysed multimodal and multisensory data of the trainees debate behaviour, and contrasted it with those of skilled professional debaters. Observational, correlation and machine learning studies were performed to identify multimodal markers of persuasive speech and link them to experts' assessments. A combination of multimodal in- and out-of-domain debate data, and various non-verbal, prosodic, lexical, linguistic and structural features has been computed based on our analysis and assessed used to , and several classification procedures has been applied achieving an accuracy of 0.79 on spoken debate data.
Index Terms: multimodal paralinguistics, perception of multimodal paralinguistic phenomena, multimodality in argumentative discourse

1. Introduction

Modern human-computer interactive technology is essentially multimodal. Modalities that are commonly used include speech, gestures (both "on-screen" touch and "in-air" gestures), eye gaze, haptics, etc. Multimodal dialogue is not only the most social and natural form of interaction, but is proven to have positive effects when incorporated in human learning, coaching and medical treatment or therapy [1, 2, 3].

The current state of the technology enables tracking of visible body movement and facial expressions. A huge diversity of sensors is available on the market for tracking visible movements (3D Kinect, Intel®RealSense™), eye-tracking (Tobii, SMI Glasses) and biometrical signals (Myo, Blood Volume Pulse and NeXus EXG sensors), etc.

While exhaustive real-time monitoring seems unrealistic with the current technology, and also from an ethical point of view, certain multimodal markers may be defined that trigger and guide the interaction and presentation of information. Progress has been booked in multimodal behaviour modelling, with advances in social signal processing and affective computing, see [4] for an overview. The identification of multimodal markers and their relation to psycho-physiological assessments is however still very much under development.

The use case considered in this study is the training of debate skills, which typically involves ad-hoc face-to-face classroom debates. A debater's proficiency level is often judged on three criteria: (1) argument organization, (2) argument content, and (3) argument delivery. While argument content and organization have received considerable attention of philosophers, logicians, linguists and teachers [5, 6, 7, 8], the descriptions of argument delivery and presentation are considerably less detailed and often vague. Therefore we will first consider characteristics of a 'skilled professional debate speech' with respect to linguistic, prosodic and body language features, based on previous theoretical and empirical findings (Section 2). Next, we describe our targeted domain and application, and present multimodal data collection, processing and annotations performed (Section 3). The main focus of this study is on important characteristics of political rhetoric, such as persuasiveness and assertiveness. We show that to assess a debater's assertiveness level a single indicator is often insufficient. We analysed linguistic, voice quality patterns and their correlations with co-speech gesture events, in particular beat gestures. The observed patterns are used in Section 4 to evaluate trainees' presentational performance related to their persuasive debate style.

2. Previous empirical findings: qualities of persuasive public speech

Debates (i.e. 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 attention of the audience, and express authority, confidence, respect and friendliness. People generally associate certain speech, personality and interaction features with what they think is a 'good public speaker', see e.g. [21]. 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* (words or phrases repetitions for information density reduction and emphasis, e.g. well-known 'Lists of Three' [19]), and *meta-discursive acts* to relate speaker to audience, to maintain topic-comment structure, etc. [12, 19, 13]. Prosodic and acoustic strategies in speech may be decisive in conveying an opinion in a political debate [10]. 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 [11]. Mispronounced or poorly articulated words, frequent hesitations, restarts and self-corrections negatively influence the perceived speaker confidence and may jeopardize speaker credibility [9]. Table 1 summarizes previous empirical findings on correlations observed between linguistic, acoustic and prosodic speech properties and human judgments of a 'good rhetoric'. Lexical, syntactic, and prosodic choices are not only rich and powerful communicative tools used by skilled debaters to persuade their audience, but they also influence discourse processing to a great extent (see e.g. [22, 23, 24]), while noticeable

http://dx.doi.org/10.21437/Interspeech.2017-98

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

Speech property	Correlates	
	linguistic	acoustic-prosodic
Clear articulation and fluency	disfluences, hesitations absence [9, 10] false start absence [11]	fraction of voiced/unvoiced frames frequent voice breaks
Adequate prominence and focus, topic-comment structuring	topicalization, passivization, it- and wh-cleft discourse structuring (meta-discoursive) acts [12, 13]	$>$ pitch range; $>$ mean pitch; $>$ intensity [14, 11, 15, 10, 13]
Pausing (Boundaries & Grouping)	clear syntactic structures,phrasing, chunking [13]	slowing down speech rate [16, 13] pausing [17, 18, 13]
Tempo	-	number of syllables per second
Adequate voice volume	- -	perceived as normal (60-54 dB) noticeable perceived change around 4dB
Expressiveness	$>$ repetitions (List of Three) [19] $>$ density of personal pronouns [11] $<$ information density and redundancy [12, 13] mixture of short and long sentences	variations in pitch range [20] $>$ standard deviation in pitch [11, 20]

mismatches in their production may hinder comprehension.

While syntactic, lexical and intonational patterns related to the persuasiveness of public speech are relatively well understood, its multimodal aspects deserve more attention. Effects of audio-visual prosody have been studied with a focus on co-speech gestures, in particular with respect to multimodal information status markers, such as those of focus and prominence. For example, Krahmer and Swerts (2007) investigating *visual beats* concluded that if observers see a visual beat they perceive a corresponding phrase as more prominent [25]. We may expect that prosodically prominent words and phrases when accompanied by gestures will intensify the assertiveness and persuasion effect of the debate arguments.

Brentari et al. (2013) provided a methodology to quantify the relationship between pitch accent and beat gesture events [26]. Their findings show that a gesture event coincides with or slightly precedes the co-occurring word, and its stroke partially precedes the pitch accent.

Summing up, previous research shows that, although it is often difficult to define clear properties of persuasive debate or public speech behaviour, there are certain linguistic, prosodic and body language features that correlate with human judgments of such behaviour. A cooperative conversational partner[1] makes use of these features, employing Frequency, Effort and Production Codes [27], and respecting general Conversational Maxims [28] as well as maxims related to intonational meaning [14]. Extending these prinicples to multimodal behaviour in general and to multimodal debate performance in particular, we will be not only able to explain perceptive regularities but also to formulate production rules and strategies that novice public speakers may follow to deliver convincing performance, and that can be used for its assessment:

- *authentic confidence and authority* (Frequency code - Maxim of Quality - Maxim of Pitch): 'try to match physical realization of your utterance to the degree of confidence you wish to convey', e.g. by appropriate pitch, speaking rate, verbal and non-verbal behaviour.
- *appropriate intensification* (Effort code - Maxim of Relation - Maxim of Emphasis): 'try to make informationally important portions of your speech acoustically, intonationally and visually prominent'.
- *adequate articulateness and grouping* (Production code - Maxim of Quantity & Manner - Maxim of Phrasing): 'try to phrase your speech in a way that is clear, dividing it into meaningful portions linguistically, prosodically and visually'.
- *distinguishable coherence* (Production Code - Maxim of Relation - Maxim of Range): 'try to match your linguistic and audio-visual performance to the degree of relevance of information you transfer', e.g. structure argument properly, avoid irrelevant information, increase your pitch range to start new topics.

This study focuses on a detailed analysis of multimodal markers of *confidence* and *intensification*.

3. Data: scenario, collection and annotations

This study is motivated by the design of a Virtual Debate Coach, whose main task is to train young parliamentarians how to debate successfully [29]. The system monitors the trainees' verbal, vocal performance, as well as their body posture and gestures, and provides feedback indicating what behaviours needs to be improved and how. The trainee is expected to deliver better performance and gain confidence through practicing debates (learning by doing) and through feedback from human or virtual tutors. Feedback (corrective, verification, instructional, 'try again') concerns ongoing formative assessment and summative assessment which reflects one or more debate sessions [30].

An important step in designing any multimodal dialogue system is to model natural human dialogue behaviour, based on the analysis of examples of such behaviour. Our core data collection activity involved debate *trainees*. Our target users were school children aged 14-15 years who have been exposed to very little debate training. In order to assess the trainee's performance and measure their proficiency level we make comparisons with data of *skilled* debaters who are young parliamentarians, members of the Youth Parliament, and enjoyed extensive training in a debate school or club (e.g. English Speaking Union[2]), and *professional* world-class debaters who have made a successful political career.

The collected data is referred to as the *Metalogue Debate Corpus*. It consists of 11 sessions of a total duration of appr. 2.5 hours, comprising 400 arguments (Argumentative Discourse Units) from 6 different bilingual (English/Greek) speakers. Each debate session involved a pair of participants: one of the participant is randomly assigned the role of a proposer, the other the role of an opponent. Two Kinect cameras, each facing one participant, were placed at a distance of 1.5-2m to the participants. Participants faced each other with max. 1m distance between them. Speech signals (16kHz, 16-bit, mono)

[1] Debaters may show non-cooperative behaviour towards their opponents, but they will be always cooperative towards the audience that is their actual addressee, whose information state and opinion they try to influence.

[2] http://www.esu.org/

Table 2: *Annotated gesture events distribution 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
No visible gesture event		20.50	50.7

were recorded using Tascam portable digital recorder and segmented per speaker and roughly per turn (*speaker-diarization*) manually in Audicity[3]. Participants' speech is transcribed semi-automatically by (1) running the Kaldi-based Automatic Speech Recognizer [31] and (2) correcting ASR output manually. Visual information is obtained from the data tracked by Kinect V2 sensors and contains information about all joints for hand and arm movements , i.e.frame ID, absolute time, relative time stamp and X, Y and Z coordinates. Audio, video and Kinect streams were synchronized based on absolute time stamps with frames of equal 33ms size.

Two resources were used as benchmarks[4]: *UK Youth Parliament* (UKYP)[5] debates (see also [32]) and the collection of the *American Presidency Project* (APP)[6]. The selected UKYP and APP sessions are video recorded and available on Youtube[7]. The UKYP data comprises three debate sessions with a total duration of appr. 3 hours, consists of 118 arguments from 35 different speakers, aged 11-18, addressing : (1) relationships and sex education (RSE); (2) university tuition fees and (3) young people job opportunities. The corpus is provided with automatically generated transcripts which we corrected manually and re-segmented. From APP we selected two presidential debate sessions on multiple current affairs topics between Senator Obama and Senator McCain (2008) and between President Obama and Governor Romney (2012), and one vice-presidential debate between Governor Palin and Democratic nominee Biden (2008) with a total duration of 4.5 hours. It should be noticed that the UKYP debates are mostly prepared speeches, while the Metalogue and APP debates are largely impromptu speeches.

To assess debaters' confidence level and argument clarity/fluency, our linguistic analysis was mainly focused on identification of filled pauses, i.e. stallings [33], and speech repairs [34] used by the speaker to improve an infelicitous formulation within the same turn.

As for prosodic analysis, for each frame prosodic properties were computed automatically using PRAAT [35] such as minimum, maximum, mean, and standard deviation of *pitch*, *energy*, *voicing* and speaking rate.[8]

For visual movements features we have the recorded video and the Kinect tracked data, viewed and annotated in ANVIL[9]. The co-speech gestures were annotated (beats, iconics, deictics, emblems and adaptors [36, 37]) by two independent annotators using videos featuring one person and not his partner, and without sound. A good inter-annotator agreement on average was reached in terms of Cohen's kappa of 0.64 [38]. Distribution of detected and annotated gestures events per category is provided in Table 2 in relative frequency of gesture events identified and in proportion of frames they last. It can be observed that beats are the most frequent type of gestures in an argumentative discourse and are mostly used as prominence markers.

4. Experimental design and results

We performed a series of experiments of different types, including observational studies from the collected data, Wizard-of-Oz (WoZ) experiments involving human tutors, correlation experiments measuring the strength of intensification effects, and machine-learning experiments using various training procedures and feature subsets to build predictive models for the assessment of debater confidence and intensification power.

Observational studies involved straightforward measures describing the basic features of the collected data and comparing them to those of benchmarks. We calculated feature distributions (i.e. relative frequencies) and ratios in order to identify behavioural regularities. Skilled professional debaters were observed to avoid filled pauses, editing expressions, restarts and hasty abrupt repairs. In prepared speech we have not observed any such phenomena; in impromptu professional speeches they were rather rare. Disfluencies, if they occur, have a short duration and are often phonetically similar to the following token (or its onset), which makes them less acoustically disturbing, e.g. ‘e̲h a complementary’ - ‘ə̲ ə kɒmplɪˈmɛntəri’. We also observed that professionals 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. Editing expressions that were used also have other meanings than to signal errors, hesitations or retractions (see also [39]). For instance, frequent unconscious disruptive use of editing expressions like ‘you know’, ‘I mean’, ‘kind of’ and ‘like’ does not occur at all. Skilled professional debaters use measured speaking rate, whereas the performance of trainees is less balanced in this respect. Table 3 summarizes our findings on linguistic, prosodic and temporal aspects of fluent confident speech. We report the observed lowest and upper values and do not average over speakers.

From the literature we know that prosodically prominent tokens convey important or new information. Pitch accented tokens often coincides with the focus, topic and contrast, and if accompanied by a beat gesture are perceived as even more prominent. Beat gestures are known to slightly precede a pitch accent. Our observations support these findings concerning intensity features.For instance, we observed that 95% of all manually annotated beat gestures are produced around intensity peaks, where the intensity range between the peak and its onset or offset should be greater than 4dB.

WoZ experiments were originally conducted to study the effects of human tutoring interventions and compare them with system-generated behavior, in order to evaluate system performance [40]. In this study, we used a comparable technique to measure correlations between the debate performance of trainees and the judgments of three professional debate coaches

[3]Free downloadable at http://www.audacityteam.org/

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

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

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

[7]See as example http://www.youtube.com/watch?v=g2Fg-LJHPA4

[8]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.

[9]http://www.anvil-software.org/

Table 3: *Observations summary on argument production fluency for confidence and clarity assessment of trainees vs skilled vs professional debaters and 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

Table 4: *Correlations between features and the mean confidence level value assigned by three debate coaches. (r stands for the Pearson coefficient; α indicates the maximum false positive error possible with the threshold set at .05)*

Prosodic/Acoustic features	α	r
Mean Pitch	0.370	0.047
Standard Deviation Pitch	0.000	-0.269
Min Pitch	0.633	-0.025
Max Pitch	0.000	0.318
Fraction of Unvoiced Frames (FoUF)	0.000	-0.258
Number of Voice Breaks (NoVB)	0.000	-0.356
Mean Intensity	0.000	0.262

who assigned a persuasiveness level ranging from 0 (very not-confident performance) to 5 (very confident performance). We calculated bivariate Pearson correlations to find the significance of linear relationship between the occurrence of a certain gesture and prosodic/acoustic feature and the mean confidence level assigned by the coaches, see Table 4.

Concerning the prosodic features extracted, it can be observed that standard deviation in pitch has a strong negative correlation with perceived speaker confidence: higher standard deviation is perceived as lower confidence. This is not entirely in line with the conclusions in [11], where a higher standard deviation in pitch is explained as a signal of expressiveness and positively correlating with charisma judgments, although the relation between human perception of charismatic speech may differ from those of the confident one. We found a significant positive effect of maximum pitch and explain this by the fact that confident speakers do stress important and contrastive information and speak 'up'. Similarly, significant positive effects of the mean intensity has a significant positive correlation with the confidence of speech, suggesting that confident speakers are perceived as using acoustic and intonational intensifiers. Features such as FoUF and NoVB have significant negative correlation with confidence. We found that the speaker is perceived as less confident when he or she uses a higher number of voice breaks and unvoiced frames. In sum, clear, fluent and firm speech is perceived as confident and persuasive.

Machine-learning experiments were conducted to determine whether multimodal markers are stronger predictors of confidence than uni-modal ones, since they may intensify persuasion effects. Predictive models were built using audio-visual data to train SVM classifiers, selecting different subsets of raw and normalized features described above.

We want our models to be largely language-agnostic and did not include linguistic features in training classifiers. Prosodic features are described in Section 3. Visible movement (hand motion tracked) features were extracted and computed from the Kinect output and comprise overall gesture *duration* as well computed duration for gesture stroke and retraction phases, for each hand; *handedness* for right, left or both hands movements; *X, Y, Z coordinate values* for each hand for each frame; and *X, Y, Z coordinate values* for gesture stroke and retraction phases for each hand. Prosodic and visual movements history of 5 previous frames was encoded in a feature vector. Table 5 presents the results in terms of accuracy for prosodic and visual features and their combination. The achieved accuracy of 71.19% confirms that, when performing three-class classification (confident vs not-confident vs inconclusive), multimodal markers are stronger predictors of confidence than those extracted from the speech signal or those of hand motion tracked. Prosody remains powerful when it comes to more fine-grained decisions as shown in five-class classification (very confident vs rather confident vs rather not-confident vs very not-confident vs inconclusive or neutral). All built classifiers outperform the majority class (neutral) baseline of 39%. Additionally, our observation shows that frame-based classification, while allowing to track the smallest changes in prosody and motion is probably not the most suitable method when it comes to relate these changes to human judgments. We need to relate to verbal elements to make picture complete. For this, we will explore token-based approaches in the future. Despite current limitations, the trained classifiers turned out to be extremely useful in obtaining new annotated data, reducing annotation costs significantly (appr. 40-50% in terms of annotation time). Prediction models were used to pre-annotate debate data, which were subsequently converted to Anvil format, and edited using this tool.

Table 5: *Classification results in terms of accuracy obtained on different type of computed features.*differs significantly from the baseline according to two-sided t-test, $t < .05$*

Feature type	Accuracy (in %)	
	Three-class problem	Five-class problem
Hand motion	61.59*	41.42*
Prosody	67.54*	50.12*
Motion + prosody	71.19*	48.01*

5. Conclusions and future research

In this paper we described possible multimodal markers and their relations to perceptive properties of debate performance. In line with previous empirical findings, we acknowledge that persuasive speech is rather difficult to characterize. Nevertheless, based on theoretical and empirical frameworks set up by Grice (1975), Gussenhoven (2002) and Hirschberg (2002), we were able to define a set of criteria which help us to explain observed regularities and define rules, strategies and constraints for the generation, assessment and correction of trainees' debate performance. Experiments of different types supported fairly reliable identification of markers from multimodal data, and linking these to assessments of debater confidence level and intensification behaviour.

We intend to continue this study in the future in two directions. First, we will incorporate our findings in the Virtual Debate Coach, enabling the system to automatically detect and interpret variations in debate behaviour, assess debater proficiency level, and provide feedback aiming at an *immersive* user experience. Pilot experiments with users indicated that we are on the right track, see [40, 29]. Second, we will incorporate more sophisticated lexical, syntactic, semantic and pragmatic features to discover new regularities, constraints and relations.

6. References

[1] S. Sali, N. Wardrip-Fruin, S. Dow, M. Mateas, S. Kurniawan, A. A. Reed, and R. Liu, "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, 2010, pp. 179–186.

[2] B. Woods, E. Aguirre, A. E. Spector, and M. Orrell, "Cognitive stimulation to improve cognitive functioning in people with dementia," *Cochrane Database Syst Rev*, vol. 2, no. 2, 2012.

[3] T. F. Hughes, J. D. Flatt, B. Fu, C.-C. H. Chang, and M. Ganguli, "Engagement in social activities and progression from mild to severe cognitive impairment: the myhat study," *International psychogeriatrics*, vol. 25, no. 04, pp. 587–595, 2013.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

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

[6] D. N. Walton, *Argumentation schemes for presumptive reasoning*. Routledge, 1996.

[7] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed, "Automatic detection of arguments in legal texts," in *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ser. ICAIL '07. ACM, 2007, pp. 225–230.

[8] A. Zohar and F. Nemet, "Fostering students' knowledge and argumentation skills through dilemmas in human genetics," *Journal of Research in Science Teaching*, vol. 39(1), pp. 35–62, 2002.

[9] C. Tuppen, "Dimensions of communicator credibility: An oblique solution," *Speech Monographs*, vol. 41:3, pp. 253–260, 1974.

[10] D. Braga and M. A. Marques, "The pragmatics of prosodic features in the political debate," in *Speech Prosody 2004, International Conference*, 2004.

[11] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, 2009.

[12] R. Nir, "Electoral rhetoric in israel - the television debates. a study in political discourse," *Language Learning*, vol. 38:2, p. 187=208, 1988.

[13] P. Touati, "Temporal profiles and tonal configurations in french political speech," *Working Papers in Linguistics*, vol. 38, pp. 205–219, 2009.

[14] J. Hirschberg, "The pragmatics of intonational meaning," in *Speech Prosody 2002, International Conference*, 2002.

[15] A. Pejčić, "Intonational characteristics of persuasiveness in serbian and english political debates," *Nouveaux Cahiers de Linguistique Française*, pp. 141–151, 2014.

[16] E. Strangert, "Phonetic characteristics of professional news reading," *PERILUS XII*, pp. 39–42, 1991.

[17] A. Wichmann, "Attitudinal intonation and the inferential process," in *Speech Prosody 2002, International Conference*, 2002.

[18] E. Strangert, "Prosody in public speech: analyses of a news announcement and a political interview." in *INTERSPEECH*, 2005, pp. 3401–3404.

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

[20] P. Touati, "Prosodic aspects of political rhetoric," in *ESCA Workshop on Prosody*, 1993.

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

[22] D. Dahan, M. K. Tanenhaus, and C. G. Chambers, "Accent and reference resolution in spoken-language comprehension," *Journal of Memory and Language*, vol. 47, no. 2, pp. 292–314, 2002.

[23] D. G. Watson, M. K. Tanenhaus, and C. A. Gunlogson, "Interpreting pitch accents in online comprehension: H* vs. l+ h," *Cognitive Science*, vol. 32, no. 7, pp. 1232–1244, 2008.

[24] S. Repp and H. Drenhaus, "Intonation influences processing and recall of left-dislocation sentences by indicating topic vs. focus status of dislocated referent," *Language, Cognition and Neuroscience*, vol. 30, no. 3, pp. 324–346, 2015.

[25] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, 2007.

[26] D. Brentari, G. Marotta, I. Margherita, and A. Ott, "The interaction of pitch accent and gesture production in italian and english," *Studi e Saggi Linguistici*, vol. 51, no. 1, pp. 83–101, 2013.

[27] C. Gussenhoven, "Intonation and interpretation: Phonetics and phonology," in *Speech Prosody 2002, International Conference*, 2002.

[28] H. P. Grice, "Logic and conversation," *1975*, pp. 41–58, 1975.

[29] J. V. Helvert, V. Petukhova, C. Stevens, H. de Weerd, D. Børner, P. V. Rosmalen, J. Alexandersson, and N. Taatgen, "Observing, coaching and reflecting: Metalogue - a multi-modal tutoring system with metacognitive abilities," *EAI Endorsed Transactions on Future Intelligent Educational Environments*, vol. 16, no. 6, 2016.

[30] E. H. Mory, "Feedback research revisited," *Handbook of research on educational communications*, pp. 745 –784, 2004.

[31] D. Povey, "The kaldi speech recognition toolkit," in *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[32] V. Petukhova, A. Malchanau, and H. Bunt, "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. Suarez, Eds. Springer, Berlin, 2016, pp. 26–46.

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

[34] J. Besser, "A corpus-based approach to the classifiation and correction of disfluencies in spontaneous speech," 2006, master Thesis, Saarland University, Saarland, Germany.

[35] P. Boersma and D. Weenink, "Praat: doing phonetics by computer. computer program," 2009, available at http://www.praat.org/.

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

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

[38] J. Cohen, "A coefficient of agreement for nominal scales," *Education and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[39] V. Petukhova and H. Bunt, "Towards a multidimensional semantics of discourse markers in spoken dialogue," in *Proceedings of the Eighth International Conference on Computational Semantics (IWCS)*, H. Bunt, V. Petukhova, and W. S., Eds., Tilburg, 2009, pp. 157–168.

[40] A. Malchanau, V. Petukhova, H. Bunt, and D. Klakow, "Multidimensional dialogue management for tutoring systems," in *Proceedings of the 7th Language and Technology Conference (LTC 2015)*, Poznan, Poland, 2015, pp. 482–486.