Prosodic Plot of Dialogues: A Conceptual Framework to Trace Speakers' Role

Vered Silber-Varod[1(✉)], Anat Lerner[2(✉)], and Oliver Jokisch[3(✉)]

[1] Open Media and Information Lab,
The Open University of Israel, Ra'anana, Israel
vereds@openu.ac.il
[2] Department of Mathematics and Computer Sciences,
The Open University of Israel, Ra'anana, Israel
anat@cs.openu.ac.il
[3] Institute of Communications Engineering,
Leipzig University of Telecommunications (HfTL), Leipzig, Germany
jokisch@hft-leipzig.de

Abstract. In this paper we present a proof-of-concept study which aims to model a conceptual framework to analyze structures of dialogues. We demonstrate our approach on a specific research question – how speaker's role is realized along the dialogue? To this end, we use a unified set of Map Task dialogues that are unique in the sense that each speaker participated twice – once as a follower and once as a leader, with the same interlocutor playing the other role. This pairwise setting enables to compare prosodic differences in three facets: Role, Speaker, and Session. For this POC, we analyze a basic set of prosodic features: Talk proportions, pitch, and intensity. To create comparable methodological framework for dialogues, we created three plots of the three prosodic features, in ten equal sized intervals along the session. We used a simple distance measure between the resulting ten-dimensional vectors of each facet for each feature. The prosodic plots of these dialogues reveal the interactions and common behaviour across each facet, on the one hand, and allow to trace potential locations of extreme prosodic values, suggesting pivot points of each facet, on the other.

Keywords: Social context · Role · Positioning · Dialogue-Games
Talk proportions · Pitch variations · Intensity · Distance measure

1 Introduction

Discourse analysis studies examine the way speakers project their identity [1], and their social characteristics, via content analysis. This paper aims to merge two domains – prosodic analytics and discourse studies – by tracing the footprints of the extra-linguistic information of the role one plays in spoken interactions. Positioning is a term mentioned by [2] to reflect a conversational phenomenon, defined as the process whereby speakers' selves are located as perceptibly and subjectively coherent participants in jointly produced conversations. In speaking and acting from a position, a person is bringing to a particular situation his/her history as a subjective being – that is

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 636–645, 2018.
https://doi.org/10.1007/978-3-319-99579-3_65

the history of multiple positions and engagements in different forms of discourse that a person has been in. According to [2], the term positioning reflects the dynamic aspects of an interaction in contrast to the way in which the use of the term role serves to emphasize static and formal aspects. In practice, studies in this field deal with how the voice of the individual, in its metaphoric sense, via discursive means, [3], reflects the way participants are locating themselves in certain contexts, for example, in institutional discourse [4].

Automatic role identification methods were developed in [6–10], as part of the field of Spoken Language Understanding (SLU) system [5]. These studies were mainly focused on automatic identification of the roles exhibited by different speakers, for the purpose of automatic speaker diarization, which serves as a mechanism to attribute the automatic speech recognition output to the relevant speaker [11–14]. All these studies were varied in terms of discourse types and languages, but on the other hand, they only dealt with a single-speaker-per-role problem. The challenge of power relations modeling, on the other hand, is both a consequence of a certain (formal) role acquired by the speaker (e.g., the host in broadcast talk), and the subjective positioning of the participants (e.g., the prime minister as an interviewee). Moreover, power relations are dynamic, and may change over the duration of the interaction. A few studies have been dealt with automatic tagging of the dominant speaker. [15] found that the top five of most discriminative features in a meeting are: number of times a speaker grabs the floor, number of turns, number of successful interruptions, amount of words spoken, and number of questions asked. In previous studies, [16, 17] found evidence of prosodic-acoustic discriminative role cues of the same speaker who played different roles in two different dialogues. The findings showed a mean of 71% correct role classification rate for women and a mean of 76% classification rate for men, based on machine-learning algorithms with 1,428 acoustic features that were extracted via openSMILE [18]. As opposed to studies that rely on rich set of acoustic features, [19] found an indication to discriminative durations of silent pauses by each of the examined roles. Such findings are indicating how prosodic cues are used by speakers to manage their own communication skills, and how the role affects primitive prosodic characteristics. Other studies used surface parameters such as the above mentioned: turn number, turn duration, and turn-taking to measure conversation structures and its effect on likability [20], and other studies reached 91% accuracy using prosodic features, among these are speaking length and energy, to cue dominance and subordination relations [5].

In this proof-of-concept (POC) study, we follow the studies that showed how prosodic feature can discriminate between speakers' power relations. By using the pairwise settings of Map Task dialogues mentioned in [17], we present a conceptual framework that can be used to trace dominance and subordination relations between speakers and to compare between the same speaker's behavior when s/he plays different roles. We expect that beyond the roles speakers are assigned to at the beginning of the session, there will also be positioning processes along the interaction, and power relations between the two interlocutors. In the context of the task-oriented dialogues, power is asymmetrical due to the knowledge the leader has, i.e., the full route on the map, as opposed to the "blindness" of the follower. On the other hand, in these

task-oriented dialogues, there is also solidarity, since both sides want to fulfill the task, meaning they are in socially equal relations.

One of the theory-driven questions in sociolinguistics is whether the leader sounds like a leader, assuming leaders sound less hesitant, more restrained, and with a certain amount of charisma, as expected from a person who holds the knowledge and authority [21]. This is in contrast to the follower, whom is expected to be more hesitant and anxious, as expected from a person who is guided and does not hold the full information. With these assumptions in mind, the aim of the study is to find acoustic cues to dominance and subordination relations [15] between the speakers and to identify the vocal patterns and acoustic cues that reveal the inequality between the two roles.

2 Materials and Methods

2.1 Map-Task Corpus Pairwise Setting

This POC focuses on the speech signal of interlocutors in a task oriented dyadic interaction, namely Map Task corpus [22, 23], in which a speaker takes over either the role of a follower or a leader. To this end, we use the Hebrew Map Task corpus (MaTaCOp) [24], in which each of the thirty–two speakers participated twice with the same interlocutor – once as a follower and once as a leader. This pairwise setting allows to compare, first, the speaker's vocal characteristics in both roles, and second, the dynamics of dominance and subordination relations between the speakers along the session, as a function of the role. Map Task dialogue type of discourse is considered task–oriented, unplanned spontaneous speech [25], in which, participants have no a–priori knowledge about the recordings' setting or material (mainly, maps). According to the pairwise setting, we denoted the first session of each pair as A and the second session as B. We further denoted the speaker that began as follower (F) in session A and changed into leader (L) in session B with FL and the other speaker with LF.

In the current study, we demonstrate our preliminary results on three pairs of speakers (out of sixteen pairs): Male and male (sessions 1A and 1B), female and male (sessions 4A and 4B), and female and female (sessions 8A and 8B).

The comparisons were made in three facets:

1. *Role* – we examine how the same role (follower or leader) is played by different speakers and different sessions.
2. *Speaker* – we examine how the same speaker (LF or FL) is playing two different roles in two different sessions.
3. *Session* – we examine differences in the same session (first session A, or second session B), between two different speakers in two different roles.

Overall, gender differences are also relevant to the topic of subordination and dominance, but they will not be discussed in the current paper.

2.2 Design

As a pre-process, we conducted a single-layer automatic annotation of the dialogues. Then we parsed each dialogue into chunks and calculated talk proportions, mean pitch (F0), and mean intensity (in dB). Lastly, we calculated the relevant Euclidian Distances (ED).

Annotation Method. A single-layer automatic annotation was applied. Annotation included detection of a minimal set of four speech tags that are beyond annotators' agreement: speech units of each role (Leader or Follower), acoustic silences, and overlaps [16, 17]. The segmentations and annotations were automatically converted to PRAAT textgrids [26].

Plot Comparable Framework. Our first goal is to represent speakers' participation level along the dialogue. Following [15] who found that the amount of speech is both subjective and objective evidence to dominancy; we claim that this feature might also represent the dominancy level of speakers along the dialogue. Calculations are rather simple – measures of the amount of speech per frame (i.e., time unit). Frames can be either a fixed predefined time (i.e. a minute) or a relative unit such as a tenth of the total dialogue length. Figure 1 demonstrates three sequential Map Task sessions, 1A and 1B (top), 4A and 4B (mid), and 8A and 8B (bottom). The same two speakers participated in 1A and 1B, another pair of speakers in 4A and 4B, and yet another pair in 8A and 8B. Each of these six dialogues was divided into equal units of 1/10 of the dialogue length, to enable a comparison between the six sub-figures.

This visualization highlights the differences between these dialogues, in terms of speaker's participation level, even when the same task (albeit different maps) and the same pair of speakers are involved. The plot framework can be used to compare not only participation level, but also other prosodic components such as the speech energy of each speaker in the two different roles, or between two speakers in the same session – and the same for pitch. Thus, with the same design, we will want to measure the changes over time, and interactions between the speakers, and between the prosodic components.

Feature Extraction. For this POC study, we chose speech duration, pitch and intensity data to demonstrate the feasibility of our conceptual framework. Pitch and intensity are primitive features that have perceptual impact on listeners and best accuracy rates in automatic classifiers of conversational intelligence systems (inter alia, [27]). The feature extraction was carried out after converting the stereo sound file into its two mono channels (The recording setup is presented in [17]). For each channel, the relevant speaker's intervals were automatically extracted via Praat [26]. For example, "4A follower" intervals were extracted in channel 1 since the follower in that session had channel 1's mono microphone on him. For each of the tenth frames, we extracted the pitch (in Hz) and intensity (in dB) mean values using Praat version 6.0.32 [26]. The normalized pitch was calculated for each frame as the mean value of the frame divided by the global F0 mean value of a speaker (in a specific session, A or B). The normalized intensity was calculated for each frame with the reference intensity of the global mean intensity value of a speaker (in a specific session, A or B).

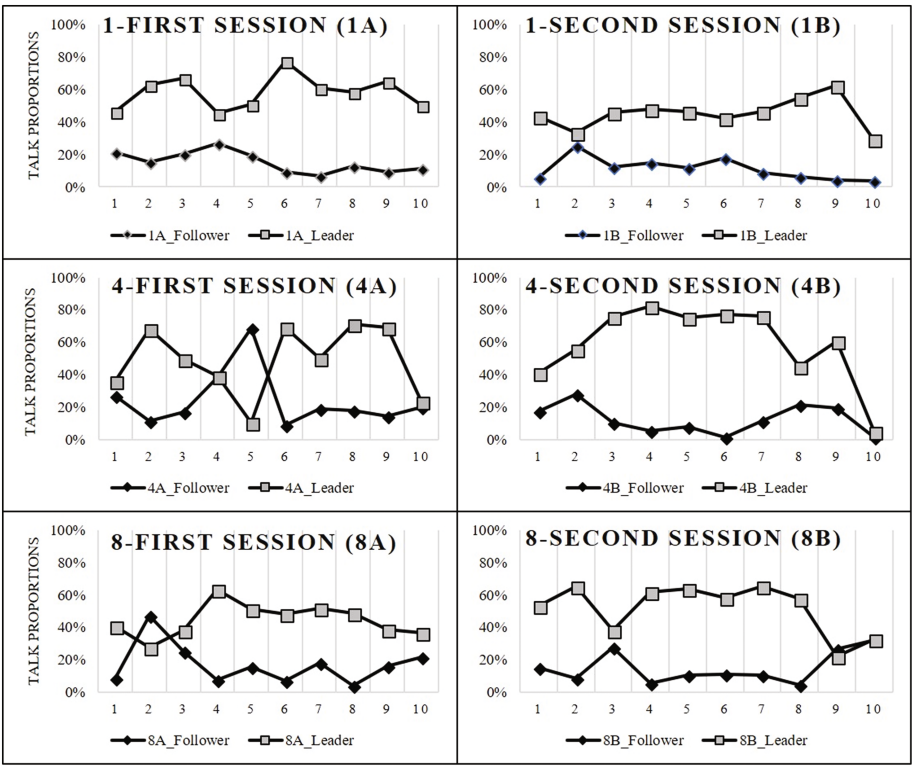

Fig. 1. A comparison between dialogues' plots of the three pairs of sessions (1, 4, and 8) according to the talk proportions (%) per each frame (1/10 of total dialogue duration).

Distance Measure. Calculations of the distance were done according to the simple ED measure for each of the above-mentioned comparison facets: Role, Speaker, and Session. For example, to calculate the distance between two followers, let $(\bar{A}_\imath)$ denotes the mean of the ıth tenth for the follower in session A, and $(\bar{B}_\imath)$ denotes the mean of the ıth tenth for the follower in session B, then the ED between the two followers is:

$$\sum_{i=1}^{10} \left(\bar{A}_\imath - \bar{B}_\imath\right)^2 \tag{1}$$

For each pair of speakers and each prosodic parameter there are six EDs: two per each facet: Role, Speaker and Session. To estimate the magnitude of the distances, we compared them to the mean distance of the six relevant EDs.

3 Preliminary Results

In the following we present the results from three different pairs of speakers that participated in the MaTaCOp recordings. The prosodic plots of the dialogues show the normalized values of the three prosodic parameters in each tenth of a session for each facet: role, speaker, and session. For each prosodic parameter we calculated the two EDs per facet and compared it to the mean ED of all six pairs (two pairs per each of the three facets).

Regarding distances in the Role facet, findings show that in terms of talk proportions, leaders differ more than followers, since their ED is higher than that of the followers (Table 1). However, the values of the mean EDs are even higher (1.00 in the pair of sessions #1, 1.23 in the pair of sessions #4, and 0.94 in the pair of sessions #8), therefore these distances between the talk proportions of two speakers who play the same role are considered low.

Table 1. Euclidean distances of talk proportions per frame, between the roles – followers and leaders.

Session	Role	Euclidean distance (ED)	Mean ED
1	Followers	0.28	1.00
1	Leaders	0.57	
4	Followers	0.76	1.23
4	Leaders	0.94	
8	Followers	0.43	0.94
8	Leaders	0.49	

Table 2 presents the EDs for pitch. The followers' EDs in sessions 1 and 8 are strictly higher (marked with boldface) than the mean ED values and in session 4 it is almost equal to it.

Table 2. Euclidean distances of normalized pitch between the roles – followers and leaders.

Session	Role	Euclidean distance (ED)	Mean ED
1	Followers	**0.42**	0.28
1	Leaders	0.18	
4	Followers	0.30	0.32
4	Leaders	0.19	
8	Followers	**0.27**	0.24
8	Leaders	0.18	

Table 3 shows that in terms of intensity, all the EDs between the followers are higher (marked with boldface) than the mean, while all the leaders are lower than the mean ED.

Table 3. Euclidean distances of normalized intensity between the roles – followers and leaders.

Session	Role	Euclidean distance (ED)	Mean ED
1	Followers	**7.89**	7.39
1	Leaders	4.82	
4	Followers	**14.57**	10.44
4	Leaders	6.33	
8	Followers	**11.38**	7.62
8	Leaders	3.85	

Regarding distances in the speaker facet, we found that for the talk proportion parameter, comparisons within speakers (FL and LF) have higher ED values than their mean values (except for speaker 1LF). ED values of pitch and intensity are higher than the mean ED in sessions 4 and 8. This tendency demonstrates how speakers play each role differently.

Regarding distances within a given session, results show that in terms of talk proportions, the EDs between the speakers in each session are above the mean. This finding is not surprising as it reflects two different speakers in two different roles. However, for the pitch and intensity, the results are not conclusive, similar to the results of the speaker facet. Figure 2 illustrates how the three prosodic parameters might be integrated to a unified prosodic plot in future research. The dynamics between the two speakers/roles in session 4A is evident. For example, the upward talking trend of the follower in the 4th and 5th frames is accompanied with higher pitch values of the follower and lower intensity values of the leader.

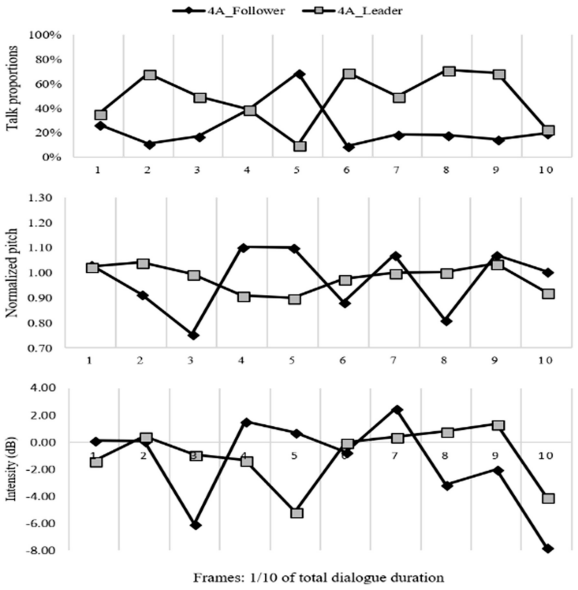

Fig. 2. The three prosodic parameters of session 4A as a typical example.

4 Discussion

In this paper we presented how speakers' participation level can be the primitive infrastructure to any type of spoken dialogue analysis. By measuring the talk proportions of each speaker in a conversation, as well as changes in pitch and intensity, a comparable plot of the dialogues might be achieved. These graphs can be of use, for example, to trace power relations between speakers.

Regarding the Role facet, we found that followers differ more than leaders in terms of pitch and intensity. This might be related to the fact that leaders' role is defined and straight forwards, while the followers' role is to seek the unknown and therefore each speaker finds her/his own way to manipulate the voice in order to fulfill the task. Since this phenomenon repeated in all three sessions, we find it interesting to further explore in this direction in large scale corpus.

Regarding the Speaker facet, we showed that speakers manipulate their voice and level of participation according to the role they were assigned to.

Regarding the Session facet, we found high distances within session. Moreover, if we think about the first sessions (A-sessions) as the first part of the dynamics between two speakers, then we should expect an entrainment effect to occur, and the difference in pitch and intensity between speakers to shrink in the second sessions, hence EDs to be higher in A-sessions compared to B-sessions. However, we did not find this trend. In future work, we intend to examine the effect of entrainment phenomenon [28] and to integrate it into our model.

This conceptual framework can contribute to a new perspective of speech detection and recognition technologies, which are classically designed to reach a real-time performance, for example, [27]. By plotting the dialogues in this manner, potential locations of extreme prosodic values are emerging. Our conceptual framework suggests analysing first the durational structure, and then to drill down into recognizing lower level features, which are derived of the initial analysis. In the present paper we showed how dialogue structure can imply pitch and intensity anomalies. Furthermore, using this analysis process, the linguistic content can be then contextualized. Therefore, we intend to widen the scope of the investigation into linguistic and perceptual cues of dominance and subordination interactions between speakers. Beyond, this POC aims to contribute to automatic role recognition [11, 16, 17, 29, 30]; to human-computer interface sciences; to speaker verification [31]; and to the emerging business field of Conversation Intelligence (CI).

Acknowledgements. This work was supported by the Open Media and Information Lab at The Open University of Israel [Grant Number 20184].

References

1. Jenkins, R.: Social Identity, 4th edn. Routledge, London and New York (2014)
2. Davies, B., Harré, R.: Positioning: the discursive production of selves. J. Theor. Soc. Behav. **20**(1), 43–63 (1990)

3. Kupferberg, I., Green, D.: Troubled Talk: Metaphorical Negotiation in Problem Discourse. Mouton de Gruyter, Berlin (2005)
4. Heritage, J., Clayman, S.: Talk in Action: Interactions Identities and Institutions. Wiley Online Library, Oxford (2010). https://doi.org/10.1002/9781444318135
5. Tur, G., De Mori, R.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Wiley, New York (2011)
6. Hori, C., Hori, T., Watanabe, S., Hershey, J.R.: Context-sensitive and role-dependent spoken language understanding using bidirectional and attention LSTMs. In: INTERSPEECH, pp. 3236–3240 (2016)
7. Ma, W., Zhang, M., Liu, Y., Ma, S.: Multi-grained role labeling based on multi-modality information for real customer service telephone conversation. In: Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), New York, pp. 1816–1822 (2016)
8. Chen, P.C., Chi, T.C., Su, S.Y., Chen, Y.N.: Dynamic Time-Aware Attention to Speaker Roles and Contexts for Spoken Language Understanding. arXiv preprint arXiv:1710.00165 (2017)
9. Chi, T.C., Chen, P.C., Su, S.Y., Chen, Y.N.: Speaker Role Contextual Modeling for Language Understanding and Dialogue Policy Learning. arXiv preprint arXiv:1710.00164 (2017)
10. Li, Y., et al.: Unsupervised classification of speaker roles in multi-participant conversational speech. Comput. Speech Lang. **42**, 81–99 (2017)
11. Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S.: The rules behind roles: identifying speaker role in radio broadcasts. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pp. 679–684 (2000)
12. Liu, Y.: Initial study on automatic identification of speaker role in broadcast news speech. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pp. 81–84. Association for Computational Linguistics (2006)
13. Weizman, E.: Positioning in Media Dialogue: Negotiating Roles in the News Interview. John Benjamins Publishing, Amsterdam/Philadelphia (2008)
14. Zhang, B., Hutchinson, B., Wu, W., Ostendorf, M.: Extracting phrase patterns with minimum redundancy for unsupervised speaker role classification. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 717–720 (2010)
15. Rienks, R., Heylen, D.: Dominance detection in meetings using easily obtainable features. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 76–86. Springer, Heidelberg (2006). https://doi.org/10.1007/11677482_7
16. Lerner, A., Silber-Varod, V., Batista, F., Moniz, H.: In search of the role's footprints in client-therapist dialogues. In: Proceedings of Speech Prosody 2016 (SP2016), Boston, USA (2016)
17. Silber-Varod, V., Lerner, A., Jokisch, O.: Automatic speaker's role classification with a bottom-up acoustic feature selection. In: Proceeding of the GLU 2017 International Workshop on Grounding Language Understanding, pp. 52–56 (2017). https://doi.org/10.21437/glu.2017-11
18. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010). https://doi.org/10.1145/1873951.1874246
19. Silber-Varod, V., Lerner, A.: Analysis of silences in unbalanced dialogues: the effect of genre and role. In: Eklund, R., Rose, R. (eds.) Proceedings of DiSS 2017, The 8th Workshop on Disfluency in Spontaneous Speech, MH-QPSR, Stockholm, Sweden, vol. 58, no. 1, pp. 53–56 (2017)

20. Weiss, B., Schoenenberg, K.: Conversational structures affecting auditory likeability. In: Proceedings of the INTERSPEECH, pp. 1791–1795 (2014)
21. Biadsy, F., Rosenberg, A., Carlson, R., Hirschberg, J., Strangert, E.: A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech. In: Proceedings of the Speech Prosody, Campinas, Brazil, pp. 579–582 (2008)
22. Anderson, H., et al.: The HCRC map task corpus. Lang. Speech **34**(4), 351–366 (1991)
23. Carletta, J., Isard, A., Kowtko, J., Doherty-Sneddon, G.: HCRC dialogue structure coding manual. Human Communication Research Centre (1996). http://www.lancaster.ac.uk/fass/projects/eagles/maptask.htm
24. The Map Task Corpus of the Open University of Israel (MaTaCOp). http://www.openu.ac.il/en/academicstudies/matacop/pages/default.aspx
25. Ochs, E.: Planned and Unplanned Discourse. In: Syntax and Semantics: Vol. 12. Discourse and Syntax. Academic Press, New York (1979)
26. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program]. Version 6.0.35. http://www.praat.org/. Accessed 16 Oct 2017
27. Walther, M., Neuber, B., Jokisch, O., Mellouli, T.: Towards a conversational expert system for rhetorical and vocal quality assessment in call center talks. In: Proceedings of the 6th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2015), Leipzig, pp. 29–34, September 2015
28. Pardo, J.S.: On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. **119**(4), 2382–2393 (2006)
29. Salamin, H., Vinciarelli, A., Truong, K., Mohammadi, G.: Automatic role recognition based on conversational and prosodic behaviour. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 847–850 (2010)
30. Dufour, R., Estève, Y., Deléglise, P.: Characterizing and detecting spontaneous speech: application to speaker role recognition. Speech Commun. **56**, 1–18 (2014)
31. Park, S.J., Yeung, G., Kreiman, J., Keating P.A., Alwan, A.: Using voice quality features to improve short-utterance, text-independent speaker verification systems. In: Proceedings of INTERSPEECH 2017, pp. 1522–1526 (2017)