In search of the role's footprints in client-therapist dialogues

Anat Lerner [1], *Vered Silber-Varod*[2], *Fernando Batista*[3]*and Helena Moniz*[4]

[1] Mathematics and Computer Science Department, The Open University of Israel

[2] The Research Center for Innovation in Learning Technologies, The Open University of Israel

[3] L2F – INESC-ID, and ISCTE-IUL, Instituto Universitário de Lisboa, Portugal

[4]L2F – INESC-ID, and FLUL/CLUL, Universidade de Lisboa, Portugal

`anat@cs.openu.ac.il, vereds@openu.ac.il,`
`fernando.batista@inesc-id.pt, helenam@l2f.inesc-id.pt`

Abstract

The goal of this research is to identify speaker's role via machine learning of broad acoustic parameters, in order to understand how an occupation, or a role, affects voice characteristics. The examined corpus consists of recordings taken under the same psychological paradigm (Process Work). Four interns were involved in four genuine client-therapist treatment sessions, where each individual had to train her therapeutic skills on her colleague that, in her turn, participated as a client. This uniform setting provided a unique opportunity to examine how role affects speaker's prosody. By a collection of machine learning algorithms, we tested automatic classification of the role across sessions. Results based on the acoustic properties show high classification rates, suggesting that there are discriminative acoustic features of speaker's role, as either a therapist or a client.

Index Terms: Role identification, client-therapist dialogue, speech analysis, acoustic features, machine learning.

1. Introduction

The essence of social communication relies not only on the lexical representation of the message, but also on other linguistics representation such as prosodic patterns and non-propositional content, such as paralinguistic information. Moreover, while paralinguistic cues in the voice signal convey rich dynamic information about a speaker's intentions and emotional state, extralinguistic cues are said to "reflect more stable speaker characteristics including identity, biological sex and social gender, socioeconomic or regional background, and age" [1]. An interesting extra-linguistic cue is the organizational status of a speaker, or one's personal status, compared to their interlocutor. It seems only natural and intuitive that speakers manipulate their speech and language characteristics according to the conversational *role* they exhibit. For example, one may be at the same time a mother to a child and a daughter to her elderly parents. Alternatively, one can be a supervisor of a student, and at the same time subordinate to a boss. This extra-linguistic cue of role affects the speaker's intentions and attitudes (i.e., the paralinguistic aspects), and consequently, the corresponding linguistic representations carried out by speech.

In this paper we pose the question: does the role in therapeutic sessions (i.e., therapist or client) is marked by acoustic and prosodic parameters? We hypothesize that therapist speech differ considerably from the client speech, even for a given speaker undertaking different roles. We are thus providing acoustic evidences for the discrimination of the role of the speaker in a therapeutic context.

We examine this question in a unique corpus in which each speaker practiced genuinely the two different roles in two different therapeutic sessions. This research setting allows us to examine how speakers change their prosodic use according to the role they practice. As far as we know, no previous research on role recognition was carried out on the same speaker's natural role manipulations.

The remainder of this paper is structured as follows. Section 2 overviews the related work. Section 3 describes the data collection and the annotation process. Section 4 presents our experiments and results achieved. Section 5 discusses our findings and broader implications of the study. Section 6 presents the conclusions and future work.

2. Related work

Work on the effect of speaker's occupation on her/his language has emerged from a variety of disciplines. Sociolinguistics is one of the core areas where occupational styles are investigated [2]. Beyond occupational styles, sociolinguistic studies have been contributing to the understanding of distinct biological (gender and age of the speakers) and social variables (professional status, geographic birth place, inter alia), which have been correlated with very informative acoustic-prosodic feature sets. [3] showed that a speaker uses clearly different voice qualities when talking to different interlocutors. In a language acquisition perspective, it was found that phoneme type development of infants depends on the occupational status of their parents [4]. Clinical work uses occupational speech to plan relevant interventions for individual clients, and in speech disorder research to better describe participant characteristics [5]. Speech perception studies also showed how familiar occupational phraseology contributes to intelligibility [6]. Within the faculty of speaker recognition, voice quality research is specifically concerned with the amount of different voices one can produce [7], and with the way speakers project their identity, and their social characteristics.

Prosodic research on dialogues in general had gained considerable attention (see [8], for a summary on prosodic research on dyadic speech). However, dialogues in a clinical

context only recently have been studied in an acoustic-prosodic perspective (e.g., [9]).

Characterizing the prosody of client-therapist sessions is in fact studying features of a specific discourse community and its various genres. Client-therapist sessions are unique in the sense that recording them is a challenging task, since a researcher should seek to provide an open atmosphere in order to achieve real authenticity and naturalness. Works were done in the framework of varied physiological and psychological states of the clients, and in varied ages of clients, from children to adults. Within the realm of related works on prosodic characteristics of in-session discourse are, inter alia, [10], who studied nonverbal features in the psychotherapeutic process in order to find interactional indicators which predict psychotherapy to be successful or not. An example of acoustical analysis of therapy session's recordings can be found in [11]. They found that features derived from the formant and power spectral density measurements are the best discriminators of three groups of subjects: normal individuals, individuals carrying diagnoses of depression, and high-risk, near-term suicidality. [12] summarized research on expressions of emotions in analogue therapy session in Hebrew spoken corpus. The results of these studies converged to show that specific speech and acoustic parameters can be employed to identify and distinguish between (unresolved) anger and sadness. In this context [13] introduced Emotional Prosody Measurement (EPM) as a tool to measure the intensity of emotion in order to learn about the therapeutic effectiveness. [13] also pointed out the advantages of acoustic measurements of the speech signal on other direct physiological measures, such as EMG or EEG.

To summarize, in the present study we examine the prosody of speaker's role in a genuine dyadic interaction, in order to improve our understanding of the relationship between acoustic-prosodic parameters and the social-professional dynamics. Given the session unbalanced formal structure, where therapist is a responder while the client is the principle reporter, this task is challenging. Yet, this study's framework is interdisciplinary in its nature (prosody and speech processing) and not an application-oriented research, therefore we see this as a Proof of Concept study, which will lead to the examination of more fine-tuning dialogue phenomena.

3. Data and Method

3.1 Corpus

The corpus analyzed in this study consists of audio and video recordings collected during genuine training sessions of Process Work psychological paradigm [14], [15], and [16]. The discourse type is considered unscripted and natural, according to the definition from [17].

The recordings consist of four sessions, in which four Process Work interns (women) participated. One of the interns was a client; another was a therapist and the other two were observers. Although the corpus is based on interns' sessions, the participants were not college students, but four professionals in the peak of their careers, and who were deeply involved in the process and with one another. In fact, these sessions were part of a peer-group of these four women who met regularly at least once a month for about six years before the recordings were carried out. Therefore, we believe that the presence of two women as observers had a minor effect on the naturalness of their colleague client, as well as their colleague therapist, in each session. The women's age ranged from 30 to 67, all four had also practice record of at least two years beyond the internship scope. Moreover, each of the women had her own profession (two hold a senior faculty position, one holds a management position in the high tech industry, and one conducts personal coaching and mentoring).

Each session consists of 22-26.5 minutes. This duration is similar to the dialogues' duration in [18], who used six dialogues of 20 minutes each, and similar in size to varied emotional databases, mentioned in [17].

All four participants took part in the experiment on a voluntary basis, with consent for the audio/video recording of the experimental session and the use of these recordings for scientific research. The specific goal of the experiments was revealed neither before nor after the sessions. In the following, the women will be referred to as B, G, S, and M, after the first letter of their names, and each session will be named with a two-letters abbreviation: BM, MS, SG, and GB, where the first letter is after the therapist's name and the second letter after the client's name.

In the present paper, only the audio recordings are analyzed. The data is based on high-level audio recordings (Zoom-H4n recorder, sampling rate of 44100 Hz, and the audio data is encoded as 16-bit linear PCM in little-endian convention). The design of this study has control over the factors (i) gender (only women), (ii) occupational age (working adults, above 30), (iii) language (Hebrew), (iv) no diagnosed psychological disorder, (v) the setting (same day, same room, same seating arrangements), and (vi) type of interaction (only internship therapeutic sessions).

3.2 Annotation

Emotionally charged speech (as typically occurs in psychotherapy) calls for "utterances" as the unit of analysis [12], yet since the present research seeks to rely on acoustic information only, annotation was carried out according to clear vocalic "events". Thus, the unit of analysis are inter-pausal units. We annotated four categories: client's speech, therapist's speech, overlaps between the two speakers, and silent pauses (a period of acoustic silence longer than 100 ms). In cases of overlap with other background noises (for example, traffic, children in the play yard, etc.), the speech intervals were ignored. Segmentation in the present research was carried manually by PRAAT software [19].

Although the sessions were taken under specific psychological paradigm, they are varied in terms of these basic parameters, such as the relative amount of speech for each interlocutor (and the derived amount of silent pauses). In the present study, only the two speech parameters (clients' and therapists'), which consist between 40%-66% of a session, are examined, since the focus is on the classification of prosodic features that are extracted from the speech signal.

3.3 Feature Extraction

Each audio file was segmented and labeled manually according to the speaker and her role in the session (either therapist or client). Based on the manual segmentation described above, the open source OpenSMILE feature extractor ([20] and [21]) has been used for extracting the features. OpenSMILE is capable of extracting a very wide range of speech features, and has been successfully applied in

a number of paralinguistic classification tasks (see list of references in [22]). Our experiments use a feature vector of 722 features, corresponding to a subset of the features used in the Interspeech 2010 Paralinguistic challenge on emotion detection. A complete description of the features can be found on [23]. Still, few adjustments were taken: 1) we are using a vector of 12 MFCCs instead of 14; 2) functionals were restricted to Extremes, Regression and Moments, not including Percentiles and Times. In summary, our experiments use 36 low level descriptors (LLD) + 36 delta regression coefficients times 10 functionals, and also pitch onsets, and turn durations.

4. Experiment and Results

The experiment aimed to measure the accuracy rate of automatic machine learning role classification, by merging the features of all four clients versus all four therapists. We then analyze the most important acoustic features that were involved in the role classification.

Experiments here described use Weka [24], a collection of machine learning algorithms. Different methods have been tested, including C4.5 decision trees (J48 in Weka) and Support Vector Machines (SMO in Weka). Results were achieved using 10-fold cross-validation, using the standard parameters of each one of the methods. Results are evaluated in terms of accuracy (proportion of correctly classified instances over the number of instances) and Kappa statistic (henceforth Kappa). Kappa measures the agreement between predicted and observed classes, and evaluates how strong the model is [25]. That is, a value close to 0 means that the classification is poor and that could be done almost by chance, while a value close to 1 corresponds to a very adequate model.

The first test was carried out in order to learn if the role a person conducts could be classified by acoustic parameters. The correct classification of this role demonstrates high correct rates ranging from 78% to 90% (Table 1). The results were achieved by down-sampling to 122 instances per role in each session, in order not to bias the results.

In order to further examine the role discrimination effect, regardless of the speaker identification, we conducted classification test for all four therapists' features versus all four clients' features. In this test, we ignored the identification of the speakers. The correct classification ratio is: 71%, Kappa statistics: 0.421. That is, there is a significant difference between the roles, even if all four clients and all four therapists are processed together.

Table 1. *Correctly classified roles per speaker and for all four speakers.*

Speaker	#Client instances	#Therapist instances	Correct	Kappa
M	166	159	78.3%	0.566
S	203	366	83.6%	0.672
G	147	453	90.6%	0.812
B	254	122	88.5%	0.771
M+S +G+B	763	1107	71.0%	0.421

For the merged files of the four speakers, an analysis of the most relevant features, which were involved in the classification process, was carried out. An attribute selection based on the WEKA [24] attribute evaluator Info Gain was used, which sorts the most relevant features for the task. The relevance of the first features for the role classification is presented in Table 2, with ranking, which represent a hierarchy of importance.

Table 2. *The relevance of each attribute to the role classification to all the sessions in general.*

	Attribute
1	mfcc_sma[5]_amean
2	F0final_sma_kurtosis
3	lspFreq_sma[0]_linregc2
4	logMelFreqBand_sma[1]_amean
5	lspFreq_sma_de[3]_linregerrA
6	lspFreq_sma[0]_amean
7	voicingFinalUnclipped_sma_skewness
8	voicingFinalUnclipped_sma_amean
9	logMelFreqBand_sma[4]_stddev
10	F0final_sma_de_kurtosis
11	logMelFreqBand_sma[5]_stddev
12	F0final_sma_de_skewness
13	lspFreq_sma[1]_amean
14	lspFreq_sma_de[3]_stddev
15	lspFreq_sma_de[3]_linregerrQ
16	jitterDDP_sma_de_linregc2
17	logMelFreqBand_sma[2]_amean
18	F0final_sma_skewness
19	mfcc_sma[0]_stddev
20	mfcc_sma[0]_amean

The dominant feature are those that are widely recognized as being crucial to perform speech discrimination tasks, such as gender, age, speaker diarization, paralinguistic events detection, etc. In all these tasks, features such as MFCC, PCM are very important features, almost transversal. Since a role one's takes is a non-linguistic parameter, as gender, age, etc., that might be a reason for this feature dominance. Yet, other features associated with stress, such as jitter and standard deviation of f0 or f0 final are among these top 20 most important features.

We then measured the dynamics of the speaker's speech during the session, as a function of her role. We divided each session into three equal parts, and named them: "Early", "middle", and "late". The results that are presented in Figure 1 demonstrate the possible acoustic differentiation between three stages (albeit arbitrary) of the session as a function of each of the roles. Meaning, each of the interlocutors, as a client or as a therapist, has observational dynamic along the session. Moreover, for three speakers (B, G, and M), results show higher correct percentage for the client role (black histograms in Figure 1, Kappa statistic range between 0.31-0.66) than for the therapist's role (white histograms in Figure 1, Kappa statistic range between 0.36-0.43). Speaker S is more varied along the session as a therapist than as a client. According to the results speaker M is considered the least varied along the session (lowest classification percentages), both as a therapist (49%, Kappa 0.18) and as a client (53%, Kappa 0.24).
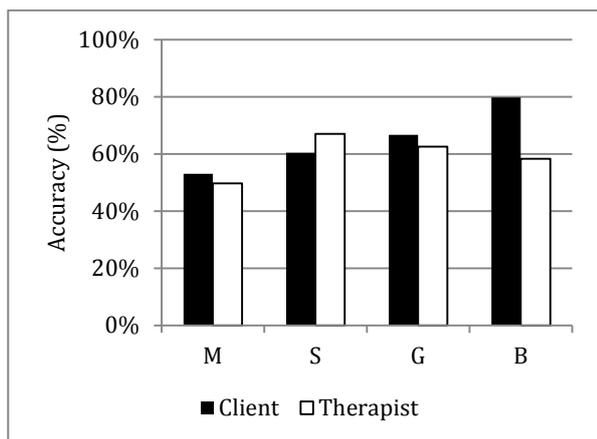
Figure 1 *Classification of the three parts (early-middle-late) between roles of the same speaker.*

Further to the above results, we have calculated how well we could distinguish between the merged client and therapist test in different parts of the session. We have balanced the data using the WEKA Filter "SpreadSubsample", and limiting the number of samples to 183 (parameters: -M 1.0 -X 183 -S 1). Results show that classification between the two roles differs throughout the session (Figure 2), and it seems that classification between therapist and client becomes more and more difficult.
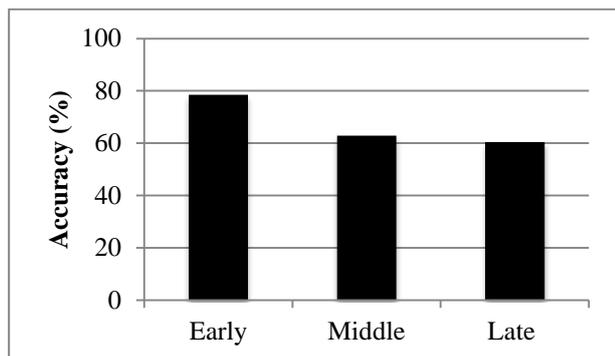

Figure 2. *Correctly classified roles (%) in three parts dynamic: Early, middle, and late.*

5. Discussion

In this paper we focused on discriminative role features of client and therapist speech interaction along a therapeutic session. Role footprints were found across tests, i.e., role is marked by acoustic and prosodic parameters. Even when facing evidence about the four speakers' divergence, we can conclude that therapist speech differs considerably from the client's speech, thus providing acoustic evidences for the discrimination of the role of the speaker in a therapeutic context. The inter-roles analysis aimed to measure whether the role of a speaker can be discriminated. We found acoustic differences between the roles for each speaker, even when all four clients and all four therapists were processed together. We showed that the role is classified with 71% correct rates.

The most relevant features in the role classification are characterized by features associated with stress and anxiety, as well as those that are widely recognized as being crucial to perform speech discrimination tasks (From these results, one can learn also about the ecology of the recordings, i.e.,

authenticity and genuine). These features should be further analyzed with respect to the dynamic of the dialogues, since we showed that role identification is different along the session (at least for the three arbitrary equal parts that we investigated).

We believe the results of this proof of concept research are promising for the study of role detection. With minimal, yet basic, annotation and massive amount of acoustic data we succeeded in demonstrating several interactional behaviors between clients and therapists. In future research, we intend to widen this type of corpus, where each speaker is engaged in two roles in two different dialogues, and to explore this phenomenon with the textual content, in order to deepen these insights. We also intend to focus on the two other dialogical parameters: pauses and overlaps, as predictive features to the role identification (and consequently, to the temporal context of the dialogue), as pauses function as discourse markers for turn taking and are a direct speech act that a speaker operates in order to pass the turn from himself to his interlocutor [26] and [27].

6. Conclusion

The contributions of the present research are twofold. First, we formed a corpus for role identification in which each participant participates twice: once as a therapist and once as a client. Second, our findings show acoustic properties discriminative of the role of the speaker as either a therapist or a client, confirming our initial research question and contributing to the discussing on acoustic-prosodic properties of role effects in speech.

This study is a step forward in what regards the scarce literature on acoustic-prosodic analysis of dyadic interaction in general, and therapeutic sessions in particular.

Regarding methodological procedures, our research suggests a methodology to trace the footprint of the role one is playing. The method is based on minimal set of objective speech tags that is beyond annotators' agreement (speakers, acoustic silences, and overlaps), and with no time consuming transcriptions. Methods as we suggest can further contribute to automatic role recognition [28], [29], and [30], to client-therapist automatic assessment tools, such as The Motivational Interviewing Skills Code (MISC) [31], and to human-computer interface [32].

We showed that with this minimal, yet precise, tagging, we can learn about the common characteristics of the different roles (as opposed to experimental prosody research that focuses on specific tokens within dialogues (e.g., [32]). Yet, being aware of the limited quantity of speech (~2h), which is a result of the ethical challenges for collecting such data, we regard this as a Proof of Concept (POS) study and further research is due.

7. Acknowledgements

8. References

[1] S. R. Schweinberger, H. Kawahara, A. P. Simpson, V. G. Skuk, and R. Zäske "Speaker perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.

[2] R. Ide, and T. Terada, "The historical origins of Japanese women's speech: from the secluded worlds of" court ladies" and" play ladies," *International journal of the sociology of language*, vol. 129, no. 1, pp. 139–156, 1998.

[3] N. Campbell, and P. Mokhtari. "Voice quality: the 4th prosodic dimension," In *15th ICPhS*, pp. 2417-2420, 2003.

[4] O. C. Irwin, "Infant speech: the effect of family occupational status and of age on use of sound types," *Journal of Speech & Hearing Disorders*, vol. 13, pp. 224-226, 1948.

[5] C. Baylor, K. Yorkston, T. Eadie, R. Miller, and D. Amtmann, "Levels of Speech Usage: A self-report scale for describing how people use speech," *Journal of medical speech-language pathology*, vol. 16, no. 4, pp. 191, 2008.

[6] C. A. Simpson "Occupational experience with a specific phraseology: Group differences in intelligibility for synthesized and human speech," *The Journal of the Acoustical Society of America*, vol. 58, no. S1, pp. S57–S57, 1975.

[7] J. Kreiman, and D. Sidtis, *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Chichester, United Kingdom: Wiley-Blackwell, 2011.

[8] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1, pp. 31–43, 2002.

[9] C. C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[10] U. Altmann, R. Hermkes, and L. M. Alisch, "Analysis of nonverbal involvement in dyadic interactions," In *Verbal and Nonverbal Communication Behaviours* (pp. 37–50). Germany: Springer-Verlag Berlin Heidelberg, 2007.

[11] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *Biomedical Engineering, IEEE Transactions on*, vol. 47, no. 7, pp. 829–837, 2000.

[12] D. Rochman, and O. Amir, "Examining in-session expressions of emotions with speech/vocal acoustic measures: An introductory guide," *Psychotherapy Research*, vol. 23, no. 4, pp. 1–13, 2013.

[13] E. L. Van Den Broek, "Emotional Prosody Measurement (EPM): A voice-based evaluation method for psychological therapy effectiveness," *Studies in Health Technology and Informatics*, pp. 118–125, 2004.

[14] A. Mindell, *The Dreambody in Relationships*. New York and London: Penguin, 1987.

[15] A. Mindell, *Dreambody: The body's role in revealing the self*. Portland, Oregon, USA: Deep Democracy Exchange, 2011.

[16] J. Diamond, and L. S. Jones, *A Path Made by Walking: Process Work in Practice*. Lao Tse Press, 2004.

[17] E. Douglas-Cowie, N., Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech communication*, vol. 40, no. 1, pp. 33–60, 2003.

[18] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," *INTERSPEECH 2009: 10th annual conference of the international speech communication association*. Baixas: isca-inst Speech Commun. Assoc, 2009.

[19] P. Boersma, and D. Weenink, Praat: doing phonetics by computer [Computer program]. version 5.3.60, 2013. Online: http://www.praat.org/, accessed 8 December 2013.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," *ACM International conference on Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[21] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," *Proceedings of the 21st ACM international conference on Multimedia*, pp. 835–838, 2013.

[22] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," *Proceedings of INTERSPEECH 2010*, Makuhari, Japan, pp. 2794–2797, 2010.

[23] B. Schuller, "The Computational Paralinguistics Challenge," *IEEE Signal Processing Magazine*. vol. 29, no. 4, pp. 97–101, 2012.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[25] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical machine learning and techniques*, 3rd edition, Burlington: Elsevier Inc., 2011.

[26] M. Heldner, and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[27] H. Moniz, F. Batista, I. Trancoso, and A. I. Mata, "Automatic structural metadata identification based on multilayer prosodic information," *Proceedings of Disfluency in Spontaneous Speech, DiSS 2013, TMH-QPSR 13*, 2013, pp. 49–52.

[28] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," *AAAI/IAAI*, 2000, pp. 679–684.

[29] H. Salamin, A. Vinciarelli, K. Truong, and G. Mohammadi, "Automatic role recognition based on conversational and prosodic behaviour," *Proceedings of the international conference on Multimedia ACM*, pp. 847–850, 2010.

[30] Y. Dufour, R., Estève, and P. Deléglise, "Characterizing and detecting spontaneous speech: Application to speaker role recognition," *Speech Communication*, vol. 56, pp. 1–18, 2014.

[31] R. Gupta, P. G. Georgiou, D. Atkins, and S. S. Narayanan, "Predicting Client's inclination towards Target Behavior Change in Motivational Interviewing and investigating the role of laughter," *Proceedings of InterSpeech 2014*, Singapore, 2014, pp. 208–212.

[32] N. Suzuki, and Y. Katagiri, "Prosodic alignment in human-computer interaction," *Toward Social Mechanisms of Android Science: COGSCI 2005 workshop*, pp. 38–44, Stresa, Italy, 2005.

[33] M. Zellers, and R. Ogden, "Exploring interactional features with prosodic patterns," *Language and Speech*, vol. 57, no. 3, pp. 285–309, 2013.