Computational modelling of speech data integration to assess interactions in B2B sales calls

Vered Silber-Varod
Open Media and Information Lab
The Open University of Israel
Ra'anana, Israel
0000-0002-1564-9350

Anat Lerner
Mathematics and Computer Science Department
The Open University of Israel
Ra'anana, Israel
0000-0002-9293-3195

Nehoray Carmi
Open Media and Information Lab
The Open University of Israel
Ra'anana, Israel
nehorayc@gmail.com

Daphna Amit
Open Media and Information Lab
The Open University of Israel
Ra'anana, Israel
amit.daphna@gmail.com

Yonathan Guttel
gong.io
Herzliya, Israel
yesguttel@gmail.com

Chris Orlob
gong.io
San Fransico, USA
chris.orlob@gong.io

Omri Allouche
gong.io
Herzliya, Israel
omri.allouche@gong.io

***Abstract*—The business sector now recognizes the value of Conversation Intelligence in understanding patterns, structures and insights of authentic conversation. Using machine learning methods, companies process massive amount of data about conversation content, vocal features and even speaker body gestures of spoken conversations. This study is a Work-in-Progress (WIP), aimed to modeling the dynamics between sales representatives and customers in business-to-business (B2B) sales calls, by relying solely on the acoustic signal. To this end, we analyze 358 sales calls at the Discovery phase. To model the conversations, we compute a basic set of acoustic features: Talk proportions, F0, intensity, harmonics-to-noise ratio (HNR), jitter, and shimmer. The plots of each acoustic feature reveal the interactions and common behavior across calls, on one hand, and within calls, on the other. The study demonstrates that using delta metrics to assess the interactions leads to new insights.**

***Keywords*-Conversation Intelligence; conversation modelling; acoustic features; speech data; sales calls.**

I. Introduction

Sales managers have always wanted to understand why some of their sales representatives consistently attain or exceed their goals while others do not [1]. In his seminal book The Tipping Point, Malcolm Gladwell wrote that a powerful personality partly means that one can draw others into her own rhythms and to dictate the terms of the interaction [2].

Recent technological advances allow automatic sales conference call recording with conversation analytics running on the back end using transcription, Conversation Intelligence (CI), and machine learning methodology. By using Artificial Intelligence to analyze information about the content, vocal features, and even body gestures, companies can understand patterns, structures and insights of sales conversations [3]. Such CI platforms aim to improve service and performance of marketing and sales personnel by providing data-driven sales conversation insights (e.g., Gong.io. [4], [5]). Orlob [4], for example, reports that talk to listen ratio in sales conversations can be used to discriminate between top and average sales representatives. Orlob [5] reports that the top, middle and bottom-ranking sales representatives listen to the customer for 54%, 32% and 28% of the conversation, respectively. Other metrics used to analyze the conversation skills of sales representatives include the ratio of customer/sales representative turn taking, the time pauses between sentences, the talk speed, and the timing, duration and content of pricing and competition discussions. Commercial companies highlight these patterns helping sales representatives to constantly improve and manage better conversations, and to efficiently stay on top of sales opportunities without the need to listen to entire conversations.

A key attribute of a conversation is its dynamics. During conversation, interlocutors must manage themselves and find within milliseconds new paths to achieve their goals. This is achieved by the Interactive Communication Management (ICM) mechanism ([6], [7]), which is comprises of features of communication that support interaction, e.g. mechanisms for management of turns, feedback, sequencing, rhythm and spatial coordination [7]. The basis of such a model, and other related theories, is the understanding that conversation processes and interactions take place within a wider social setting.

In this study, 358 authentic business-to-business (B2B) sales conversations at the Discovery phase were analyzed. The Discovery phase traditionally includes the customer describing their business history, challenges and plans ahead, followed by a demo by the sales person that attempts to relate to those topics. Discovery calls also set the trajectory for the entire deal [8].

978-1-7281-4117-6/19/$31.00 ©2019 IEEE
DOI 10.1109/DataCom.2019.00031

As the data are proprietary, we focus here on modeling the dynamics in B2B sales calls using solely the acoustic signal. Specifically, we focused on the well-known phenomenon of convergence, in accordance with the Communication Accommodation theory [9]. While the concept seems well established, as part of the study on communication, inter alia [10], and relationship-maintenance strategies [11], it has recently been studied within wider perspectives, such as neuroscience research, confirming that this phenomenon underlies successful communication [12].

Our goal is to examine the acoustic gaps between the sales representative (i.e., agent) and the customer in each conversation throughout the conversation, and to see whether there is a significant change between four sex-pairings: (1) male agent with female customer; (2) male agent with male customer; (3) female agent with male customer; (4) female agent with female customer. By focusing on the sex pairing issue, we follow recent studies that found sex pairing differences in spoken dialogues. For example, [13] found that male pairs of speakers were less efficient with respect to a map-matching task performance than female and mixed-sex pairs. Again, on a Map Task dialogues, [14] showed that there is a significant difference in the automatic role's classification rates, depending on the interlocutor's sex.

II. MATERIAL AND METHODS

A. B2B sales calls

The current dataset is based on 358 sales calls of a single company that provides a talent acquisition suite software. The calls featured two speakers, were conducted using the Zoom platform. All conversations were carried out in American English and were automatically recorded, diarized and transcribed with an in-house ASR engine (although in the current study we did not use the transcripts).

For each conversation, a JSON file was used as an input for the feature extraction process (see section B1). The JSON include time stamp of speaking utterances tagged per each speaker (either agent or customer), silence tag, and punctuation tags.

We focused on calls that are longer than 10 minutes, as many of the shorter calls were in fact unsuccessful attempts to contact the other party. The calls were conducted by 20 different sales representatives (7 females; 13 males). 103 calls were of female sales representatives (29%) and 255 calls of male sales representatives(71%). 232 (65%) calls were with female-customers and 126 with male-customers. Table I presents the distribution of mixed- and same-sex pairings.

Fig. 1 presents the number of calls per agent. Although this distribution is extremely unbalanced, it reflects the real-world characteristics of genuine data.

The dataset varies also in terms of call length. Average call duration was 44.80 minutes (Median = 50.23; SD = 18.00). Call length distribution is presented in Fig. 2 (each bin represents the maximal call length in minutes).

Table I
CALL DISTRIBUTION BY SEX PAIRING

Agent sex	Customer sex	Annotation	Amount of conversations (%)
Female	Female	FF	71 (20%)
Female	Male	FM	32 (9%)
Male	Female	MF	161 (45%)
Male	Male	MM	94 (26%)

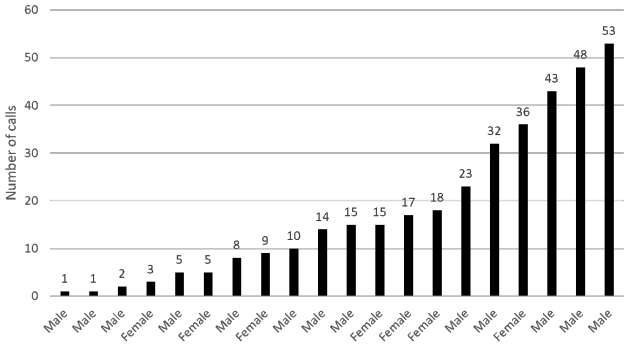

Figure 1. Call distribution per sales agent.

To examine and compare the dynamics in different interactions, we divided each interaction into 15 equal-length sections, similar to [15]. Section duration ranges from 0.6 (10 minutes call divided to 15 sections) to 6 minutes (90 minutes calls).

We then averaged each feature of every speaker in each section, obtaining 15 scores that depict the use of the feature in the conversation. We used these 15 feature scores to study the evolution (or plot) of each feature in the interaction, and to compare features expressed in a certain section to other sections.

The talk proportion model of calls longer than 10 minutes

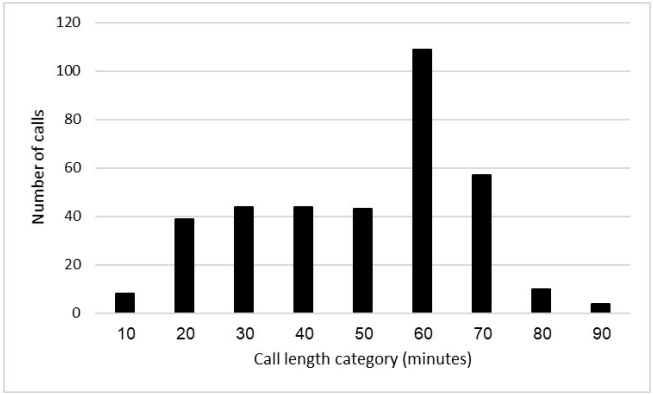

Figure 2. Call length distribution.

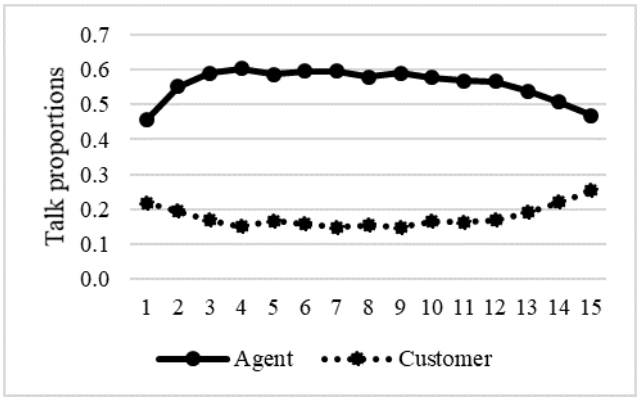

Figure 3. The average talk proportions model of calls longer than 10 minutes.

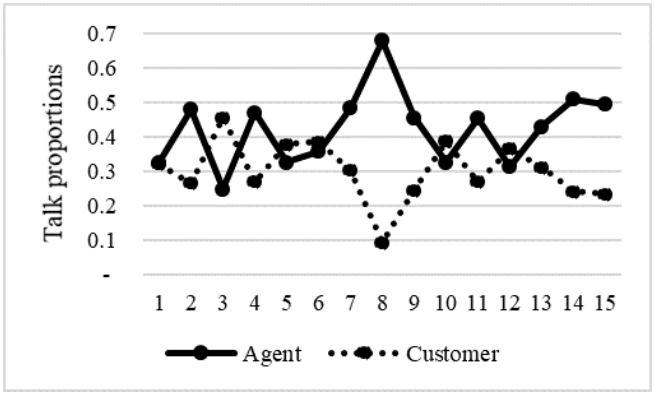

Figure 4. Talk proportions plot of short conversations (less than 10 minutes).

is presented in Fig. 3. As shown, agents talked more than customers did. However, the model for calls shorter than 10 minutes looks different, both in general (Fig. 4) and per sex pair. To understand this difference, we listened to a large proportion of such short calls and found out that participants in these calls were mainly trying to overcome communication troubleshoot and technical issues of the video conference systems. We therefore focus on calls longer than 10 minutes in the analysis.

B. Design

1) Feature extraction: For this study, we used Low-Level Descriptor (LLD) acoustic features. We extracted the mean and standard deviation of f0 (Hz), intensity (dB), HNR (through harmonicity), jitter (local) and shimmer (local) using Parselmouth protocol ([16], [17]), which is a Python interface to the internal Praat code [18]. Since the speech unit segments were utterance-based, for every speaker, we averaged each feature over a single utterance, and then averaged over the relevant section ($1/15$ of the conversation).

We further extracted the talk proportions of the two speakers, which is the ratio of the speaking time of each speaker in the given section (that is, talking time divided by section length).

2) Delta values: To represent the session's dynamics of the gap between the two interlocutors, for each session and each feature x, we use the following notations: $customer(x)$, and $agent(x)$ are 15-value vectors for feature x, a mean value per section, for the customer and agent, respectively.

V_{delta} is the main dependent variable of the current study (equation 1). To allow comparison across features, we further calculated all values with respect to the value of the first section. We denote these proportional values by V_{pdelta} (equation 2). Finally, $V_{cum-pdelta}$ is the cumulative results for the values of all the features (equation 3).

$$V_{delta}(x) = |customer(x) - agent(x)| \tag{1}$$

$$V_{pdelta}(x) = \frac{V_{delta}(x)}{V_{delta}(x)[1]} \tag{2}$$

$$V_{cum-pdelta} = \sum_{x \in features} V_{pdelta}(x) \tag{3}$$

III. Results

A. Locus of extrema delta points

We first tried to evaluate the probability that the minimum value in the vector Vdelta (Min delta) will appear after the maximum value (Max delta). The null hypothesis is that their positions are independent, and so we expect the equal probability of positive and negative differences. Fig. 5 relates to talk proportion values. It summarizes the distribution of occurrences of the Max and Min deltas in the first, the second and the third parts of the talk, respectively. We found that most of the Max deltas (about 80%) appear in the first two thirds of the calls and most of the Min deltas appear either at the beginning or at the end.

We then calculated the extrema delta points for each acoustic feature and the same trend repeated in general and in all sex pairs. Fig. 6 demonstrates the distribution of the cumulative max scores of all the seven features along the 15 sections. It is evident that in all the four sex pairs, Max delta values decrease throughout the conversation. However, different correlations were found between the sex pairs (Table II). For the distribution of the Min deltas we did not find the same, or opposite (i.e., increasing) trends (see right most column in Table II). Thus, the locus of the Max delta values, which implies a divergence between the speakers, is more predictable in our corpus than the locus of the Min deltas, which implies convergence.

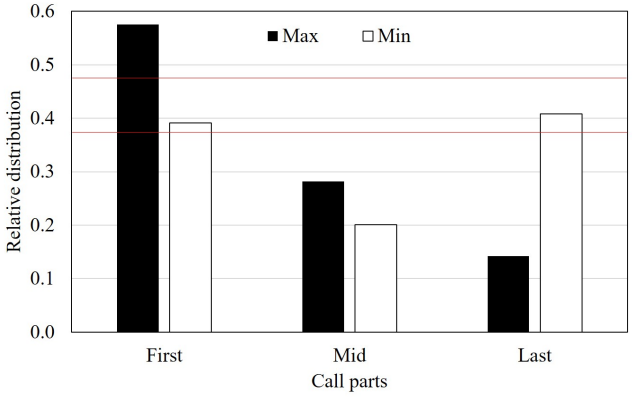

Figure 5. The relative distribution of Max and Min delta values of the talk proportions feature. The horizonal lines represent the confidence interval for 95%.

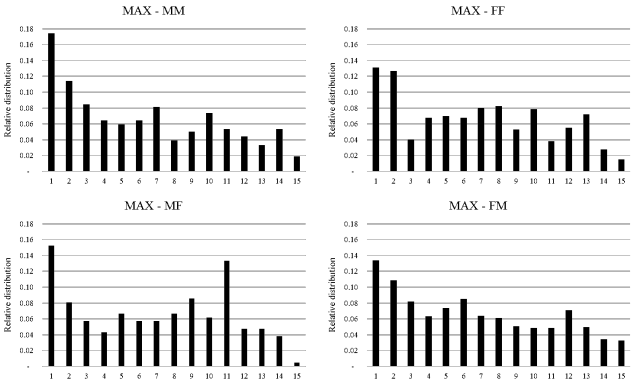

Figure 6. The relative distribution of Max delta values of all the seven acoustic features throughout the call.

Table II
CORRELATION COEFFICIENT OF MAX AND MIN DELTA DISTRIBUTION BETWEEN THE SEX PAIRS

Sex pair	Correlation of Max delta distribution	Correlation of Min delta distribution
FF-MF	0.53	0.42
MF-FM	0.58	0.63
MM-MF	0.67	0.16
MM-FF	0.76	0.27
FF-FM	0.80	0.57
MM-FM	0.87	0.17

B. The linear regression slope's measure

We then fitted the delta values with a linear model. We can say that a conversation shows a convergence process if the regression model predicts a decreasing function (negative slope), that is, smaller differences in feature delta values as the conversation progresses. For all features, we looked at the threshold proportion of calls that might show convergence for $p < 0.05$. Fig. 7 presents the differences between the four sex pairs of the convergence percentage of each acoustic feature. The horizontal lines represent the confidence interval for 95%. Out of the 28 convergence proportions (4 sex-pairs for each of the 7 features), that are presented in Fig. 7, 19 (67%) were found statistically significant. This comparison shows that convergence percentage are most varied in the FM pairs (the distance between the two red lines that represent the minimum and the maximum values for this sex pair), then in the FF pairs, followed by the MF pairs and least varied by the MM pairs (the last is also the pair with highest convergence proportions).

C. Call length variable

Theoretically, call length can affect convergence ratios since longer conversations allow more time for the interlocutors to accommodate to each other. Therefore, we examined delta values in calls shorter and longer than the median 50 minutes (not including calls shorter than 10 minutes). A one-tailed t-test showed that the difference between the call-length categories is considered to be not quite statistically significant ($p = 0.028$).

D. Cumulative delta

Per each section i, we created a cumulative mean score of the V_{pdelta}. This cumulative score combines the mean V_{pdelta} scores of all the features together. We then compared the resulted cumulative scores between the sex pairs.

The comparisons between the sex pairs is presented in Fig. 8 as cumulative flow diagrams (starting at the second section, since all V_{pdelta} scores are proportional to the first section). Significant differences were found between each of the same-sex pairs and each of the other pairs: FF and MF pairs ($p < 0.0001$); FF and FM pairs ($p < 0.05$); and FF and MM pairs ($p < 0.05$). MM and MF pairs ($p < 0.0001$); Between MM and FM pairs the difference is not quite statistically significant ($p = 0.07$); and between the mixed-sex pairs the differences is not significant.

To summarize, in terms of the cumulative scores dynamics: FF pairs are statistically different from any other pair. MM pairs are quite different from any other pair. Mixed pairs are not statistically different from each other.

IV. Discussion

In this paper, we proposed a novel measure of divergence and convergence between speakers and we applied it to analyze the B2B dialogues. The study reports evidence of convergence and a detailed comparative analysis of seven acoustic features across four sex pairs. Our study was designed to examine and to emphasize the dynamics between interlocutors in spoken conversations. We believe this data-driven innovative model, which is based on integrating diverse features, can also be used for building automated dialogue systems such as Google Duplex.

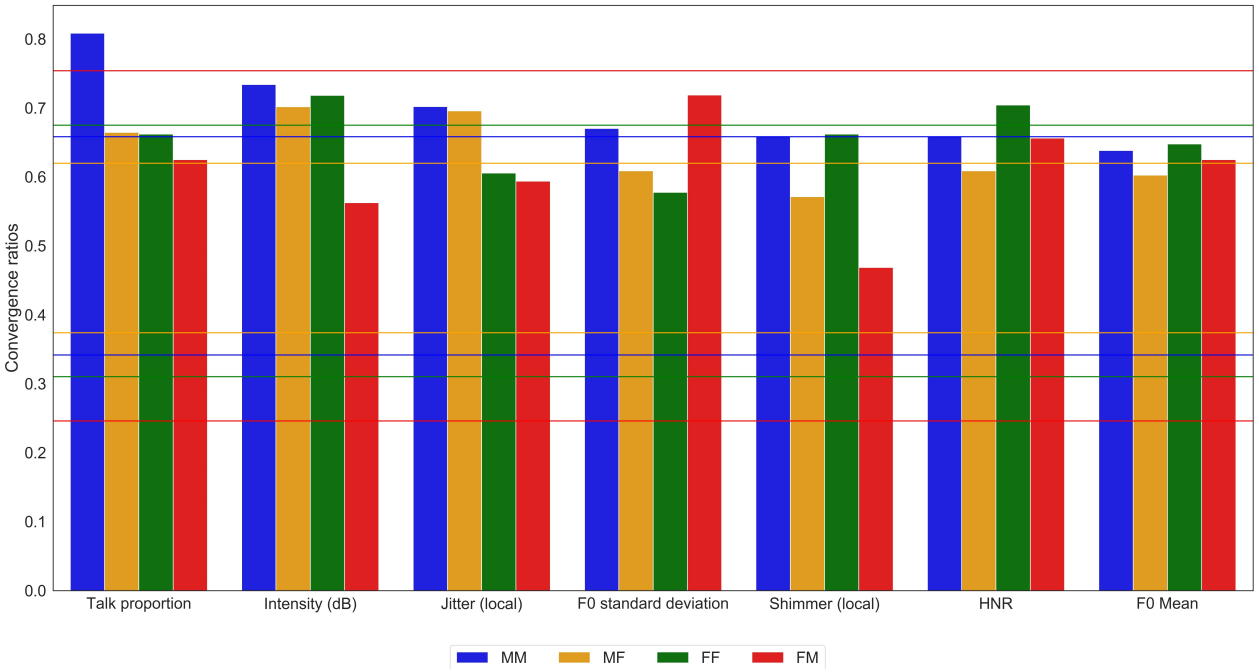

Figure 7. Convergence percentage of the seven acoustic parameters for each sex pair.

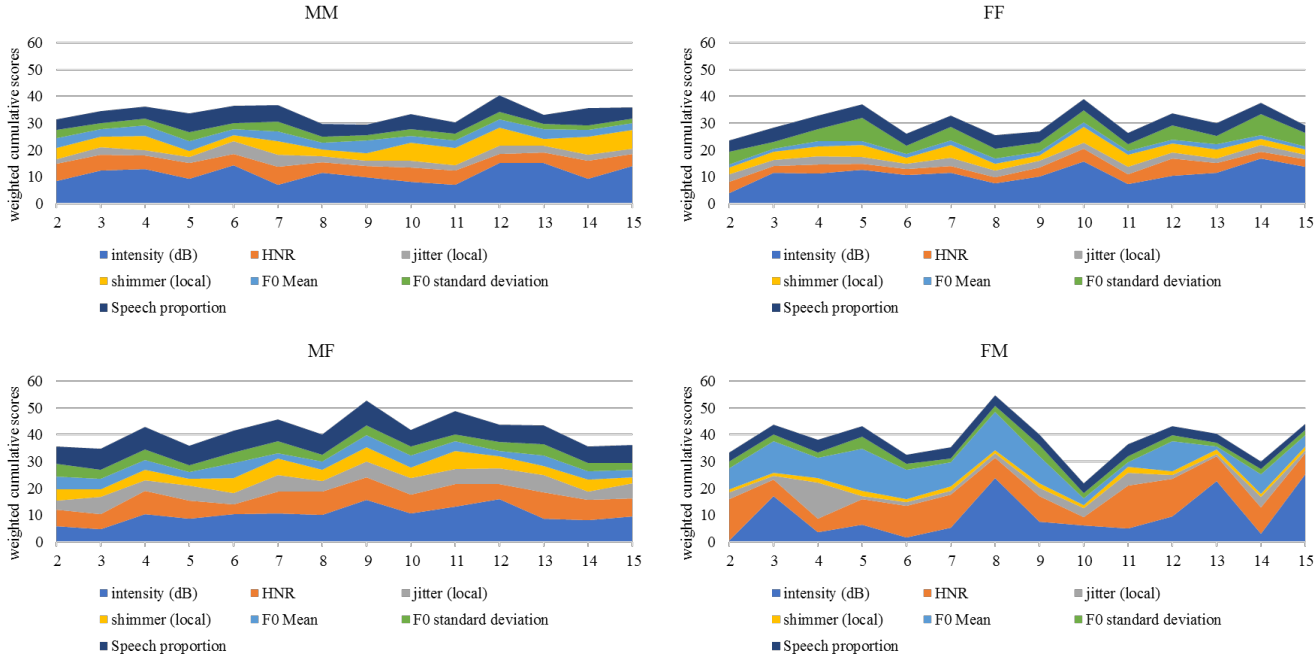

Figure 8. A cumulative diagrams of the mean delta scores of each feature in each section, per sex pair.

One of the main findings is that the loci of maximal gaps (i.e., divergence between speakers) are more predictable in this dataset, than the loci of minimal gaps (i.e., convergence between speakers). We also showed that convergence exists in all the seven features, but they are realized differently in the four sex pairs. For example, MM pairs converge more using talk proportions and intensity, while FM pairs converge more using f0 standard deviation. Moreover, findings also show that features are not necessarily in synchrony in the divergence and the convergence processes.

We also showed that acoustic convergence is not necessarily a monotonic process from the beginning to the end, and that during the conversations there are varied convergence and divergence oscillations. However, most of the talks showed convergence. As to the cumulative score, we showed that intensity gaps are prominent in all sex pairs; f0 and HNR gaps are more prominent in FM pairs; and f0 standard deviation gaps are more prominent in FF pairs. Jitter gaps are more present in mixed-sex pairs, which is interesting since this feature is known to relate to emotional state.

To conclude, the current study is a computational modelling of acoustic data integration towards assessing interactions of B2B sales calls. We demonstrated our methodology on a certain speech phenomenon – convergence – using delta metrics. In future studies we intend to apply AI and machine learning algorithm and to add discourse analysis to this integrated model.

V. ACKNOWLEDGEMENTS

This work was supported by the Open Media and Information Lab at The Open University of Israel [Grant Number 20184]. The authors wish to thank Raquel Sitman for helpful comments.

REFERENCES

[1] M. Kovac and J. Frick, "It's 10 AM. Do you know what your sales reps are doing?," Harvard Business Review, March 10, 2017.

[2] M. Gladwell, The Tipping Point: How Little Things Can Make a Big Difference, Little Brown, 2000.

[3] V. Silber-Varod, "Is human-human spoken interaction manageable? The emergence of the concept Conversation Intelligence," Online Journal of Applied Knowledge Management (OJAKM), vol. 6, no. 1, pp. 1–14, 2018.

[4] C. Orlob, "[Infographic] The science of winning sales conversations," a blog published on April 10, 2017. Available at: https://www.gong.io/blog/winning-sales-conversations/, 2017a.

[5] C. Orlob, "This is what separates your star reps from the rest of the team," a blog published on November 4, 2017. Available at: https://www.gong.io/blog/this-is-what-separates-yourstar-reps-from-the-rest-of-the-team/, (2017b).

[6] J. Allwood, "Dimensions of embodied Communication - towards a typology of embodied communication," in I. Wachsmuth, M. Lenzen, and G. Knoblich (Eds.), Embodied communication in humans and machines (pp. 1–24). Oxford, UK: Oxford University Press, 2008.

[7] J. Allwood, "The structure of dialog," in M. M. Taylor, D. G. Bouwhuis, and F. Néel (Eds.), The structure of multimodal dialogue II (pp. 3–24). Amsterdam, Netherlands: John Benjamins, 2001.

[8] C. Orlob, "7 Sales Skills You Can Dramatically Improve Using Gong.io," Gong.io Blog. September 10, 2018. Available at: https://www.gong.io/blog/sales-skills/

[9] H. Giles, "Communication accommodation theory," in B. B. Whaley and W. Samter (Eds.), Explaining communication: Contemporary theories and exemplars (pp. 293–310). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2007.

[10] R. Levitan, S. Benus, A. Gravano, & J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in 2015 AAAI spring symposium series, 2015.

[11] H. Giles, & J. Harwood, "Managing intergroup communication: Life span issues and consequences," Trends in Linguistics Studies and Monographs, vol. 100, 105–130, 1997.

[12] G. J. Stephens, L. J. Silbert, and U. Hasson, "Speaker-listener neural coupling underlies successful communication," Proceeding National Academy of Science USA, vol. 107, no. 32, pp. 14425–14430, 2010.

[13] J. S. Pardo, et al., "The Montclair Map Task: Balance, Efficacy, and Efficiency in Conversational Interaction," Language and speech, 2018. doi: 0023830918775435.

[14] A. Lerner, O. Miara, S. Malayev, V. Silber-Varod, "The Influence of the Interlocutor's Gender on the Speaker's Role Identification," In: A. Karpov, O. Jokisch, R. Potapova (Eds.). Speech and Computer (SPECOM 2018). Lecture Notes in Computer Science (pp. 321-330), vol. 11096. Springer, Cham, 2018. doi https://doi.org/10.1007/978-3-319-99579-3-34

[15] G. B. Yom-Tov, S. Ashtar, D. Altman, M. Natapov, N. Barkay, M. Westphal, and A. Rafaeli, "Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights," WWW (Companion Volume), pp. 1689–1697, 2018.

[16] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," Journal of Phonetics, vol. 71, pp. 1–15, 2018. doi.org/10.1016/j.wocn.2018.07.001

[17] https://github.com/YannickJadoul/Parselmouth

[18] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved 8 September 2018 from http://www.praat.org/, 2018.