Automatic Speaker's Role Classification with a Bottom-up Acoustic Feature Selection

Vered Silber-Varod[1], *Anat Lerner*[2], *Oliver Jokisch*[3]

[1]The Research Center for Innovation in Learning Technologies, The Open University of Israel
[2]Department of Mathematics and Computer Sciences, The Open University of Israel
[3]Institute of Communications Engineering, Leipzig University of Telecommunications (HfTL)

vereds@openu.ac.il, anat@cs.openu.ac.il, jokisch@hft-leipzig.de

Abstract

The objective of the current study is to automatically identify the role played by the speaker in a dialogue. By using machine learning procedures over acoustic feature, we wish to automatically trace the footprints of this information through the speech signal. The acoustic feature set was selected from a large statistic-based feature sets including 1,583 dimension features. The analysis is carried out on interactive dialogues of a Map Task setting. The paper first describes the methodology of choosing the 100 most effective attributes among the 1,583 features that were extracted, and then presents the classification results test of the same speaker in two different roles, and a gender-based classification. Results show an average of a 71% classification rate of the role the same speaker played, 65% for all women together and 65% for all men together.

Index Terms: role classification, acoustics, feature extraction, machine learning feature selection methods, Map Task corpus

1. Introduction

Paralinguistic cues in speech convey rich, dynamic information about a speaker's intentions and emotional state, while extra-linguistic cues are said to reflect more stable speaker characteristics, such as social identity [1], biological sex and social gender, socioeconomic or regional background, and age. Within the well-studied field of speaker recognition, voice quality research is specifically concerned with the various laryngeal degrees of freedom which result in various ways that speakers use to project their identity in ongoing interactive discourse. Parallel to such an acoustic analysis and signal processing investigations, which mainly look at how people convey non-linguistic information, scholars in the domain of discourse analysis are concerned with how speakers display their positioning in a text (i.e., the expression of one or more dimensions of the self [2], and [3]). In some studies, the term *voice* is used, in its metaphoric sense, to explain how an individual or a group locate themselves in certain contexts, for example, in institutional discourse [4]. Previous studies have shown that participants locate themselves or position themselves relative to themselves - *reflexive positioning* - and relative to their immediate interlocutors (at the time the conversation is carried out), known as *relative positioning* [5, p. 48]. Scholars claim that positioning goes beyond the immediate interaction, and that it is relative to the other people the interlocutors have been negotiating with, or in discourse with, in the past and sometimes relative to future interlocutors in a future context [6], [7], and [8]. Following this line of research, [9] studied the acoustic characteristics of the role one plays in dialogues, as part of a field of research that has flourished in recent years, namely, *prosody-in-interaction*, and is part of other linguistic and extra-linguistic phenomena that have been explored in the domain of speech-in-interaction.

Previous studies on role identification during speech-in-interaction were mainly concerned with automatic identification of the roles exhibited in broadcast news and talk shows, where automatic speaker diarization serves as a mechanism to attribute the automatic speech recognition output to the relevant speaker, for example [10], and [11]. Such studies showed how speakers were categorized into three types: anchor, journalist, and guest, based on regular expressions that each role is more likely to utter. In [9], the authors looked for evidence of prosodic-acoustic discriminative role cues in therapeutic sessions and, at the same time, studied the dynamics between the client and the therapist in the sessions. Their results, although relying on only four individuals, showed acoustic convergence tendencies between speakers in the session, and prosodic similarities for each role across sessions. Moreover, 71% correct classification rates were measured for the two different roles each of the four speakers played, indicating how an occupation, or a role, affects vocal characteristics. Following the promising results of that study, the role one plays is the extra-linguistic information under investigation in this paper.

In the current study, the *role* is an a-priori factor in these dialogues, as speakers know their role at the beginning of each session. However, the research assumes power relation processes unfolding during the interaction. One of the theory-driven questions in sociolinguistics is whether the leader sounds like a leader, assuming leaders sound less hesitant, more restrained, and exuding a certain amount of charisma, as expected from a person who holds the knowledge and authority. This contrasts with the *follower*, who is assumed to be more hesitant and anxious, as expected from a person who is guided and does not hold full information.

Concerning the prosodic realization of charisma and assertiveness, in perceptual first-impression personality judgments, [12] revealed a tendency of listeners to select the low-pitched voice over the high-pitched voice as more trustworthy for both genders, but only low-pitched male voices were significantly perceived as more dominant. Therefore, we hypothesize that leader and follower roles can be automatically identified via acoustic analysis. To this end, we used a unified set of spoken dialogues that are unique in the sense that the same speaker participated twice – once as a follower and once as a leader. This pairwise setting allowed the comparison of the speaker's vocal characteristics in both roles.

10.21437/GLU.2017-11

2. Speech data and recording setup

The current study used acoustic data from recordings of 16 Hebrew-speaking pairs of speakers (32 participants; 18 women and 14 men). The recordings were conducted according to the Map Task design [13], where one speaker (leader) instructs another speaker (follower) to reproduce a route on a map with landmarks. This type of dialogue is considered an elicited semi-spontaneous type of speech. Each participant participated twice and in two sessions: once as a leader and once as a follower (with the same partner). The corpus is called The Map Task Corpus of the Open University of Israel (MaTaCOp) [14]. In these recordings that were made from September 2015 to January 2016, we used two pairs of maps from the original set [13], with a translation of the landmarks into Hebrew.

All recordings were made according to the same setup, and the following parameters were strictly kept: distance between participants; no air-conditioning; closed windows and door; computers shut off; a carpet under the participants' chairs; the microphone close to the speaker's mouth but not touching it; comfortable adjustment of the headset ("Madonna" type) on the speaker.

The recording device was an H4N (zoom-na.com) [15]. The following parameters were chosen: activation via batteries; two paths stereo, with two passive microphones, one per speaker; 96kHz sampling rate; 24 bits; WAV format audio files; no signal processing; the 3.5 mm stereo jack microphone input split into two 3.5 mm mono jack microphone inputs in order to connect two separate mono microphones.

3. Method

In the following sections, we present the automatic segmentation method (§3.1); the feature extraction method via OpenSmile [16], and [17] (§3.2); the bottom-up method of feature selection via the WEKA tool [18] (§3.3); and the role classification process via WEKA (§3.4).

3.1 Automatic segmentation method

There are two standard approaches for segmenting the audio files into intervals for feature extraction:

Fixed regular intervals upon which the feature vector is extracted. This approach is often used for music feature extraction [19] or in a segmental feature vector normalization approach, and is carried out automatically [20].

Meaningful intervals according to the goals of the research. Sometimes this means a manual time-consuming segmentation of experts and a measure of agreement between annotators.

In this study, we chose the second option. The unit of feature extraction per each speaker is an inter-pausal unit (IPU), an *instance*. For this purpose, an automatic segmentation algorithm was written with MATLAB version 7.11.0.584 (R2010b) [21] to divide the instances from acoustic silent events and from overlapping speech (in the MaTaCOp corpus, speech instances are 75% of the duration of all sessions). The algorithm thus identified a minimal set of objective dialogue "events" that are beyond annotators' agreement (instance per speaker, acoustic silences, and overlaps), as previously mentioned in [9], and [22].

Although the sessions were made using a specific design (i.e., The Map Task), they are varied in terms of other basic parameters, such as the relative amount of speech for each interlocutor. In the present study, only the speech instances of the leaders and followers are examined, since the focus is on the classification of prosodic features that are extracted from the speech signal.

The average duration of speech instances of women speakers is 0.71 seconds, same as the average of men's. The average duration of instances of Leaders is 0.74 seconds; and of Followers is 0.65 seconds.

3.2 Feature extraction process

For feature extraction, we used OpenSmile: The INTERSPEECH 2010 Paralinguistic Challenge feature set extracted for emotion recognition [23]. We chose the IS10_paraling.conf configuration file with 1,583 attributes [23] and [24]

To these 1,583 features, we added the following two attributes – a class feature, either *Follower* or *Leader*, representing the role of the participant in each instance and the duration ($t_{max} - t_{min}$) of each instance.

3.3 Feature selection process

In order to provide more insight about the acoustic cues relevant for the role discrimination, we tried to preselect the most relevant features, although knowing that, beyond, many features from the mentioned OpenSmile set might carry additional information.

We used the following bottom up feature selection process. The process was carried out on a training set of 10 different speakers from 10 different pairs of speakers: five of which played the role of leaders in the first session (in sessions 1-5); and five played the role of follower in the first session (in sessions 6-10).

We chose the supervised class balancer of instances as a filter, since the number of instances was not balanced between the two roles. This filter reweights the instances in the data so that each class has the same total weight.

We performed a feature selection by "select attributes" with the evaluator: InfoGainAttributeEval that evaluates the worth of an attribute (feature) by measuring the information gained with respect to the class [18]. The search method ranked the attributes by their individual evaluations [18].

We selected only attributes with a worth higher than 0.1 per speaker. For each attribute, we summed its worth for all the participants in the training set, resulting in a score per attribute. We then took the 100 highest-scoring attributes as the features for classification for all speakers. This method was inspired by social welfare functions in welfare economics.

3.4 Role classification process

For the role classification, we also used the Weka tool [18]. We created csv files, a file per each participant, with all the instances related to that participant as a leader and as a follower. The class attribute was the role of the participant in each instance. For the classification phase, we used the top 100 attributes that were selected by the process described in Section 3.3 on the training set. It should be emphasized that these same 100 attributes were used for all the files in the test set. We chose a supervised class balancer of instances as a filter, since the number of instances was not balanced between the two roles. The classification was carried out via Sequential Minimal Optimization (SMO) that implements John Platt's SMO algorithm for training a support vector classifier, with cross-validation of 10 folds.

4. Results

4.1 Selected feature types

In this section, we present the feature types that were selected in the bottom-up process. As expected, the main groups of features that were effective are those that are categorized as part of the 34 low-level descriptors [16]:

- logMelFreqBand (51 different features of this type) – logarithmic power of Mel-frequency bands 0-7 (distributed over a range from 0 to 8 kHz).
- pcm_loudness (19 features) – The loudness as the normalized intensity raised to a power of 0.3.
- mfcc (19 features) – Mel-Frequency cepstral coefficients 0-14.
- lspFreq (6 features) – the 8-line spectral pair frequencies computed from 8 LPC coefficients.
- F0final (2 features) – the smoothed fundamental frequency contour.
- F0finEnv (with a single features) – the envelope of the smoothed fundamental frequency contour.
- voicingFinalUnclipped (with a single features) – the voicing probability of the final fundamental frequency candidate. *Unclipped* means that it was not set to zero when it falls below the voicing threshold.
- Duration of instances (with a single feature).

Our findings show that 51% of the attributes relate to the logarithmic power of Mel-frequency bands (logMelFreq-Band). The volume of this low-level descriptor in the OpenSmile 2010, 1,583 feature set is 21% [16]; and that 14% of the attributes consist of the outlier-robust maximum value of the contour (percentile 99.0). Moreover, all six highest scoring attributes consist of this function.

4.2 Classification tests

In the following sections, we present the classification results of two tests – the first was classifying the same speaker in two different roles, and the second, gender-based test assessed the two different roles for males and females, separately. For a baseline comparison, we first checked the classification rates for the randomly-chosen 11 speakers playing the same role. We arbitrarily marked the instances as class A or B alternately, so that all odd instances belonged to the same class, and all even instances belonged to the other class. As expected, the average classification rate for these 11 randomly-chosen sessions was: 51%, with an average kappa of 0.07. To be on the safe side, we took two different *leaders* and two different *followers* of the same gender and run their classification. As expected, we received high classification rates of 97%, with 0.95 kappa for the leaders, and 94% with 0.89 kappa for the followers. We will now present the classification of the role for 22 speakers in the test dataset.

4.2.1 Same speakers, different roles

In this test, each pair of participants was numbered from 1 to 16 (representing the 16 pairs of speakers), and a sub-notation of the role of each speaker was referred to as either a Follower (F) or a Leader (L). For each speaker, we created a file combined from all the instances of the two roles s/he played, ending with 22 files in total, representing each of the 22 speakers (Table 1). In the following, we use the nFL (and nLF) for a speaker in the pair n that participated first as F and then as L (e.g., 1LF and 1FL). Thus, nLF and nFL identify two different participants.

The classification rates as well as the kappa statistics values are presented in Table 1. All rates are above chance level. The lowest classification between two roles of the same speakers is 58.7% (found in 13LF), while the highest rate is 89.4% (found in 1FL). Interestingly, the lower rates below 70% are composed of six speakers from six different sessions, i.e., their counterparts' classification is above 70%. The average classification rate of the role per speaker is 72.8%, and 0.46 Kappa statistic

Table 1: *Role classification of 22 speakers.*

Speaker	Kappa	Classification rate (%)
13LF	0.17	58.7
10LF	0.24	61.8
6LF	0.25	62.4
15LF	0.31	65.3
9LF	0.33	66.7
11FL	0.37	68.5
12FL	0.42	71.2
5FL	0.43	71.3
4FL	0.44	72.0
11LF	0.44	72.0
13FL	0.44	72.0
14FL	0.47	73.3
12LF	0.49	74.1
7LF	0.51	75.4
16FL	0.52	75.8
14LF	0.52	76.1
15FL	0.54	76.9
16LF	0.56	78.1
8LF	0.59	79.4
3FL	0.59	79.6
2FL	0.64	82.1
1LF	0.79	89.4

4.2.2 Same gender, different roles

Finally, we tested the role classification rates for the two groups of speakers: males and females separately. We merged all male leaders and all male followers. The role classification rate was 65% with kappa 0.3; the same results were found for women leaders versus women followers.

5. Discussion

In this paper, we focused on selecting discriminative role features of Hebrew speakers who played two roles – leaders and followers – in two different sessions of a Map Task dialogues setting. Role footprints were found across all speakers, suggesting that speakers change their speech according to the role they play. The results also hint at the ecology of the setting of the present study, especially since the low classification rates were not from the same sessions and thus cannot be attributed to a session, or to the recordings, or to any other technical aspects, but rather to the speaker's speech behavior when playing the two roles.

We can conclude that leaders' speech differs considerably from followers' speech, thus providing acoustic evidences for the discrimination of the role of the speaker in a Map Task context. Moreover, we found acoustic differences between the roles, even when all male speakers and all female speakers

were processed together, and showed that the role is classified with 65% correct rates.

The most effective attributes are different to some extent than those related to the stress and anxiety that were found in the role classification of therapeutic sessions (between therapist and clients) [9]. However, similar to [9], we found attributes that are known as being crucial to perform speech discrimination tasks.

The high speaker recognition rate suggests that the 100 attributes chosen via a bottom-up method can also fit the identification of a speaker. Indeed, the challenge of finding a reduced and more limited set of features depends on many parameters [25]. Therefore, the current bottom-up approach might not be the optimal, and better results for role identification could be achieved using other approaches, and even with a top-down method. It is interesting to note that the merged role classification rate per gender was found above chance level and identical for the male combined files and the female combined files.

6. Summary and conclusions

The main contributions of this paper are threefold:

The first contribution is to automatic role recognition applications and to human computer interface. We found that speakers apply a different timbre of voice to the different roles they played (a leader or a follower in a Map Task setup). These findings further validate our previous research regarding role classification of client versus therapist in therapeutic sessions [9]. In this respect, we have demonstrated another use of the Map Task design as a corpus for role identification. In this respect, the current study can thus contribute to automatically label grounded knowledge by adding representations of the speakers' role in a dialogue. Moreover, an application for identifying the speaker's role can be of use to individuals that wish to improve their rhetoric skill, by learning how to use their voice in order to shift from a "follower" label to a "leader" labeling.

Second, regarding methodological procedures, our research suggests a bottom-up methodology to determine the effective attributes of the role one is playing, with machine-learning tools. In future research, we intend to examine the functional percentile 99.0 of voice quality parameter and logMelFreqBand. It should be noted that shimmer and jitter were not found within the top 100 attributes.

The third contribution is the database setup. To the best of our knowledge, this project is a pioneer Map Task corpus in Hebrew. We hope it will serve as a platform for rich research on Hebrew speech and language.

In future research, we intend to examine what happens when a (natural) leader play the follower and vice-versa – for example, grounding the dialogue in an institutional context of an organization, and to ask managers and their subordinates to participate in the recordings. Moreover, we would like to improve the database by adding the result, for each speaker, of a personality test. Another promising research direction will be to examine these 100 features with respect to the dynamic of the dialogues, since [25] and [9] showed that role identification rates are different at the beginning, middle, and the end of an interaction.

7. Acknowledgement

We would like to thank Jacob Azogui for his contribution to the compilation of the corpus and the feature extraction process.

8. References

[1] R. Jenkins, *Social Identity*, 4th edition, London and New York: Routledge, 2014.

[2] S. Wortham, "From good student to outcast: The emergence of a classroom identity," *Ethos*, vol. 32, no. 2, pp. 164–187, 2004.

[3] I. Kupferberg, I. Gilat, E. Dahan, and A. Doron, "Exploring the discursive positioning of a schizophrenic inpatient via method triangulation," *International Journal of Qualitative Methods*, vol. 12, no. 1, pp. 20–38, 2013.

[4] J. Heritage and S. Clayman, *Talk in action: Interactions, Identities, and Institutions*. Oxford: Wiley Online Library, 2010. DOI: 10.1002/9781444318135

[5] B. Davies and R. Harré, "Positioning: The discursive production of selves," *Journal for the theory of social behaviour*, vol. 20, no. 1, pp. 43–63, 1990.

[6] M. Bamberg, "Narrative discourse and identities," in J. C. Meister, T. Kindt, W. Schernus, and M. Stein (Eds.), *Narratology beyond literary criticism* (pp. 213–237). Berlin, Germany: Mouton de Gruyter, 2004.

[7] I. Kupferberg and D. Green, *Troubled talk: Metaphorical negotiation in problem discourse*. Berlin, Germany: Mouton de Gruyter, 2005.

[8] E. Weizman, *Positioning in media dialogue: Negotiating roles in the news interview* (Vol. 3). Amsterdam/Philadelphia: John Benjamins Publishing, 2008.

[9] A. Lerner, V. Silber-Varod, F. Batista, and H. Moniz, "In search of the role's footprints in client-therapist dialogues," *Proceedings of Speech Prosody 2016 (SP2016)*, pp. 400–405, 2016.

[10] B. Zhang, B. Hutchinson, W. Wu, and M. Ostendorf, "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.

[11] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00)*, pp. 679–684, 2000.

[12] M. Tsantani, P. Belin, and P. Mcaleer, "Low vocal pitch preference drives first impressions of trustworthiness and dominance in non- contextual scenarios," *Perception*, vol. 45, no. 8, pp. 946–963, 2016. doi: 10.1177/0301006616643675.

[13] H. Anderson, M. Bader, E. G. Bard, …, and R. Weinert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[14] J. Azogui, A. Lerner, and V. Silber-Varod, The Open University of Israel Map Task Corpus (MaTaCOp). Available at: http://www.openu.ac.il/en/academicstudies/matacop/

[15] ZOOM. ZOOM H4n handy recorder. 2016. https://www.zoom-na.com/products/field-video-recording/field-recording/zoom-h4n-handy-recorder.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the munich versatile and fast open-source audio feature extractor," *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459-1462, 2010. doi:10.1145/1873951.1874246

[17] F. Eyben, F. Weninger, M. Woellmer, B. Schuller, The Munich Versatile and Fast Open-Source Audio Feature Extractor, 2014. http://audeering.com/research/opensmile/.

[18] M. Hall, I. Witten, and E. Frank, *Data mining: Practical machine learning tools and techniques*, Third edition, Burlington: Kaufmann, 2011.

[19] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Colette, and P. Depalle, “Feature extraction and temporal segmentation of acoustic signals,” *ICMC: International Computer Music Conference*, 1998.

[20] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Communication*, *vol. 25, no.* 1, pp. 133–147, 1998.

[21] MATLAB (2010). version 7.11.0.584 (R2010b). The MathWorks, Inc., Natick, Massachusetts, United States.

[22] V. Silber-Varod, H Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir, “Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals,” *Proceedings of Speech Prosody 2016 (SP2016)*, pp. 1211–1215, 2016.

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan. “The INTERSPEECH 2010 paralinguistic challenge,” *InterSpeech, 2010*, pp. 2795-2798, 2010.

[24] F. Eyben, K. R. Scherer, B. W. Schuller, ... & K. P. Truong, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[25] 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.