Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings

Marcela Charfuelan, Marc Schröder, Ingmar Steiner

DFKI GmbH, Language Technology Lab
Saarbrücken and Berlin, Germany
firstname.lastname@dfki.de

Abstract

In this paper we investigate the prosody and voice quality of dominance in scenario meetings. We have found that in these scenarios the most dominant person tends to speak with a louder-than-average voice quality and the least dominant person with a softer-than-average voice quality. We also found that the most dominant role in the meetings is the project manager and the least dominant the marketing expert. A set of raw and composite measures of prosody and voice quality are extracted from the meeting data followed by a Principal Components Analysis (PCA) to identify the core factors predicting the associated social signal or related annotation.

Index Terms: prosody, voice quality, vocal social signals, perceptual interpretation, acoustic correlates

1. Introduction

The new research area of Social Signal Processing (SSP) is aimed at automatic understanding of social interactions through analysis of nonverbal behaviour [1]. Social signals include (dis)-agreement, empathy, hostility, politeness, dominance and any other attitude towards others that may not be expressed using just words [1]. *Vocal social signals* have to do with *how* something is said, how prosodic features like pitch, energy and rhythm, as well as voice qualities like harsh, creaky, tense, etc. are used to convey a social signal.

The present work investigates both prosody and voice quality features of social factors on real data. Recently reported works, related to the detection and classification of dominance, only use nonverbal prosodic cues and in some cases in combinations with visual and/or verbal cues. In [2], speaking length and energy as audio cues, as well as video features, are used to classify the most dominant person in a group meeting. In [3], speaking energy and speaking status, along with non-relational and relational cues derived from these vocalic cues and visual cues, are used for predicting dominance and role-based status in scenario meetings. In [4], easily obtainable features such as speaking time, number of turns in a meeting, number of words spoken in the whole meeting, etc. are used to detect dominance. Other non-verbal displays of dominance used in analysis of social interaction are presented in the exhaustive review of [5].

On the perceptual side, in the literature, dominance and status are reported to be related constructs: dominant-personality people often occupy high positions in organisations and high-status people are often allowed (even expected) to use dominant behaviour with their subordinates; the two concepts not always coincide though [3]. Vocal features such as the amount of talking time, speech loudness, speech tempo, and pitch have been demonstrated to play a role in perceptions of dominance, credibility, and leadership ability [6]. Regarding voice quality, high dominant persons have been characterised by loud, tense voice, and low dominance persons by soft, fearful voice [7].

In this paper we investigate the prosody and voice quality of dominance in the AMI scenario meetings [8] and the correlation of dominance with status or speaker role in these meetings. In contrast to previous studies, the long term goal of our investigation is not classification or detection of social signals, but automatic synthesis of expressive speech. That is, we aim to extract from real data prosody and voice quality patterns of social signals that can be used to synthesise different expressions suitable for a range of social signals. Thus, rather than building sophisticated classifiers from the extracted audio features, we must condense the findings into a form that is beneficial for synthesis. One aspect of this is to reduce the features to a human-interpretable form that allows the formulation of rules regarding the relation between the acoustic measures and the social signals.

The paper is organised as follows. In Section 2, the methodology employed to investigate and interpret the relation between acoustic measures and social signals is explained. In Section 3, the corpus and annotations used in this study are described. In Sections 4 and 5, we first analyse the acoustic correlates of speaker roles in the AMI corpus, based on the annotations of four types of speaker roles; then we compare the findings with the same analysis on a subset of the same AMI corpus for which dominance annotations are available. Conclusions and main findings of this comparison are summarised in Section 6.

2. Methodology

In order to be able to synthesise different expressions suitable for a range of vocal social signals we must (i) investigate relations between acoustic measures and vocal social signals, (ii) interpret the relation of acoustic measures and vocal social signals in view of speech synthesis, and (iii) reduce the acoustic measures to a human-interpretable form that allows us to formulate rules for re-synthesis. In this study the following procedure is applied.

1. Measures extraction: compute the acoustic measures of Section 2.1 for meeting data.
2. Clustering: use Principal Component Analysis (PCA[1]) to search for patterns, reduce redundancy among the measures and identify most relevant acoustic predictors.
3. Perceptual interpretation of acoustic measures: based on perceptual correlates reported in the literature, in Table 2, the tendencies of the acoustic measures loading most strongly on the major PCs is interpreted.

[1] PCA performed with R: http://www.r-project.org/text

2.1. Acoustic measures

The measures used in this study were extracted using Snack[2] and scripts developed in Matlab. We distinguish between raw prosodic and spectral measures and derived measures of voice quality, both of which are either computed per analysis frame or per utterance. Among the raw prosodic and spectral measures we extract fundamental frequency (F_p), energy and voicing rate, as well as frame and utterance based spectral measures related to formants (F_{np}), bandwidths (B_n) and spectrum intensity on different frequency bands (H_n, A_{np}). Frame-based measures were computed with a frame length of 25 ms and a frame shift of 5 ms.

The frame based voice quality measures presented in Table 1 are rough spectral estimates of traditional voice quality parameters normally calculated in the time domain. These measures were developed by [9] and tested successfully on the classification of emotions under different levels of noise and reverberation. They are calculated on the basis of the frame based raw measures, but some of the measures (indicated with tildes) have additionally included vocal tract influence compensation, that is, the contribution to the spectrum of each of the four formants is removed as explained in [9]. Additionally these measures are gradients instead of pure amplitude ratios, since according to their developers, gradients better characterise the shape of the glottal signal spectrum.

The utterance based voice quality measures presented in Table 1 were originally developed in [10], where various perceptual factors correlate with acoustic data from the long term average spectrum (LTAS) and fundamental frequency distribution. These measures are based on the calculation of LTAS in three frequency bands: 0-2 kHz, 2-5 kHz and 5-8 kHz. For each of these bands the maximum level is selected. These measures have been also used in emotion research [11, 12].

Table 1: *Voice quality measures. OQG: Open Quotient Gradient, GOG: Glottal Opening Gradient, SKG: Skewness Gradient, RCG: Rate of Closure Gradient, IC: Incompleteness of Closure. Tilde indicates vocal tract compensation [9], see text.*

Measure	Definition
Frame based [9]:	
OQG	$(\tilde{H}_1 - \tilde{H}_2)/F_p$
GOG	$(\tilde{H}_1 - \tilde{A}_{1p})/(F_{1p} - F_p)$
SKG	$(\tilde{H}_1 - \tilde{A}_{2p})/(F_{2p} - F_p)$
RCG	$(\tilde{H}_1 - \tilde{A}_{3p})/(F_{3p} - F_p)$
IC	B_1/F_1
Utterance based [11, 10]:	
Hamm_effort	ltas_{2-5k}
Hamm_breathy	(ltas_{0-2k} - ltas_{2-5k}) - (ltas_{2-5k} - ltas_{5-8k})
Hamm_head	(ltas_{0-2k} - ltas_{5-8k})
Hamm_coarse	(ltas_{0-2k} - ltas_{2-5k})
Hamm_unstable	(ltas_{2-5k} - ltas_{5-8k})
slope_ltas	least squared line fit of LTAS in the log-frequency domain (dB/oct).
slope_ltas1khz	least squared line fit of LTAS above 1 kHz in the log-frequency domain (dB/oct).

[2]Pitch and formants extracted with Snack: http://www.speech.kth.se/snack/

2.2. Perceptual interpretation of voice quality measures

Table 2 summarises main tendencies and correlates of the acoustic measures used in this study, with perceptions reported in the literature for different vocal social signals including dominance. A more detailed study in this respect has been presented in [13]. In this Table we can observe two main trends or tendencies of the acoustic measures: the upper group, which in general corresponds to tendencies observed in soft vocal effort, and the lower part for tendencies of loud vocal effort. These examples show that there is not a simple one-to-one mapping between perception and acoustic measures.

3. The AMI meeting corpus

The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings. Some of these meetings are naturally occurring, and some are elicited, particularly using a scenario in which the participants play different roles. This work focuses on the elicited meetings. In the scenario, four participants play the roles of employees in an electronics company that decides to develop a new type of television remote control. Although the scenario is pre-defined and the roles assigned, the conversations and discussions in the meetings reflect natural interaction [8]. This corpus contains recordings of both video and audio data, orthographic transcriptions and several levels of annotations, for example role and dominance. The transcriptions include word level segmentation time-aligned to the audio recordings.

3.1. Role and dominance annotations

For the analysis of roles, nine meeting sessions held at IDIAP were selected from the AMI corpus, corresponding to 35 sub-meetings, 36 speakers (26 male and 10 female). The audio was taken from the individual headset. The roles of the participants are: Industrial Designer (ID), Marketing Expert (ME), Project Manager (PM), and User Interface designer (UI). Role annotations for this data are available in the AMI corpus and were used to analysing acoustic correlates of speaker role.

The analysis of dominance was performed with a subset of this data for which dominance annotations are available [2]. In this case 11 sub-meetings, from five meeting sessions held at IDIAP, have been divided into 5 minute segments for which three annotators ranked the participation from highest to lowest, according to their level of perceived dominance. As described in [3], from the annotations, a significant number of the meeting segments (34) showed full agreement of the most dominant person; also there were 23 additional segments where 2 out of 3 annotators agreed on the most dominant person. In this study we have used these two sets full agreement (34 segments) and the majority agreement (34+23 segments) for analysing acoustic correlates of dominance.

3.2. Variability reduction

The acoustic measures presented in Section 2.1 were extracted for all the IDIAP meetings. Our objective is to analyse variation or patterns on the measures due to roles and dominance, but the measures are also affected by other sources of variation, including speaker gender, individual speaking style, various sources of noise including overlapping speech, outbursts such as laughter, as well as the intrinsic contextual variability which in particular includes the phonetic variability due to the uncontrolled nature of the phonetic content spoken in a natural dialogue.

Table 2: *Correlates of prosody and voice quality measures, with perception reported in the literature.*

Vocal social signal	Prosody/VQ	Acoustic measures
Affirmation and agreement [14]	short and long fall tones	
Low dominance [7, 3]	soft, fearful voice	$\downarrow F_p$, $\downarrow$SR, $\downarrow$E
Secrecy or confidentiality [15]	whisper	$\uparrow$OQG, $\uparrow$GOG, $\uparrow$SKG, $\uparrow$RCG, $\uparrow$IC,
Relaxation [16], intimacy [15]	breathy voice	$\uparrow$ Hamm_breathy, $\downarrow$ Hamm_effort
Contentness [16]	whisper	steeper slope_ltas
Intimacy [16]	lax creaky voice	flatter slope_ltas1khz
Friendliness [16]	creaky voice	
Positive politeness [17]	creaky voice	
Negative politeness [17]	sustained high pitch	$\uparrow F_p$, $\uparrow$SR, $\uparrow$E,
Surprise or unexpectedness [14]	short rise tones	$\downarrow$OQG, $\downarrow$GOG, $\downarrow$SKG, $\downarrow$RCG, $\downarrow$IC,
High dominance [7, 3]	loud, tense voice	$\downarrow$ Hamm_breathy, $\uparrow$ Hamm_effort
Admiration [14]	pressed voice	flatter slope_ltas
Happiness [11]	modal/tense voice	steeper slope_ltas1khz
Anger, stress [16]	tense/harsh voice	

When applying PCA directly on the measures per utterance, we get only very weak effects. It seems that the large amount of uncontrolled variation masks any systematic effects that may be present in the data. Therefore, it is essential to reduce the variability of the data. In a first experiment to reduce the high variability of the data, the measures were averaged per speaker, but the effects in terms of roles were still weak. An alternative to averaging was to control for phonetic content. Therefore, in a second experiment intended to reduce variability, we used only the different occurrences (tokens) of a single frequent word, "control". It would have been preferable to investigate a single vowel; however, since the AMI corpus does not contain time-aligned phonetic labels, we revert to the use of a constant phonetic segmental form. 373 one-word segments were found in the IDIAP corpus. No further averaging is applied in this reduced corpus, apart from that applied to the frame based acoustic measures.

4. Acoustic correlates of speaker roles

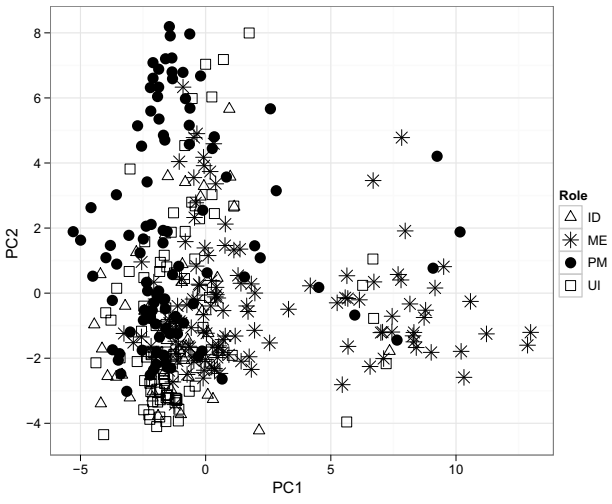

Figure 1: *Speaker roles in IDIAP meetings, first two principal components. ID: Industrial Designer, ME: Marketing Expert, PM: Project Manager, UI: User Interface designer.*

Figure 1 shows the projection of the single-word data onto a PC1-PC2 plane. Clusters for ME and PM roles are apparent that differ from the general distribution. The gender distribution of the single-word data shows that ME and PM have more female participants than UI and ID, so that any joint deviation of ME and PM could potentially be attributed to speaker gender rather than speaker role; however, it can be seen from Figure 1 that PM spreads more than average across PC2, whereas ME spreads more than average across PC1. This effect cannot be explained merely by speaker gender, but seems specific to the speaker roles.

Table 3: *IDIAP control data: proportion of variance explained by the first two PCs.*

	% Var.	Loadings
PC1	30.5	$-$GOG, $-$SKG, $-$RCG
PC2	20.3	$-$slope_ltas1khz, $+$Hamm_unstable, $+F_p$

The loadings of the PCA are presented in Table 3. The first PC accounts for 30.5% of the variance and the first 9 PCs explain more than 90% of the variance. The most loaded measures in PC1 discriminate better for ME: considering the measures GOG, SKG and RCG for ME are higher than for the other roles, this indicates a soft vocal effort tendency employed by ME. GOG for PM is smaller than for the other roles, also SKG and RCG are relatively small, especially when comparing to ME; F_p is higher than average for PM, these observations indicate a loud vocal effort tendency employed by PM. The mean value of Hamm_unstable for ME seems to be in a modal range. The slope_ltas1khz value for ME is relatively flat compared to that of the other roles, so this might also indicate a soft vocal effort tendency. Hamm_unstable and slope_ltas1khz for PM contradict the loud pattern tendency, though.

5. Acoustic correlates of dominance

Figure 2 shows the discrimination of roles and dominance according to the first two PCs on the reduced IDIAP set. The loadings for the PCA for Full most/least agreement data set are presented in Table 4. Similar results were obtained for the Majority most/least agreement data set. According to the mean values of the more loaded measures in PC1, Energy, voicing rate and F_p are higher for most dominant and lower for least dominant; on PC2 RCG, SKG, and GOG are lower for most dominant and higher for least dominant. These two tendencies

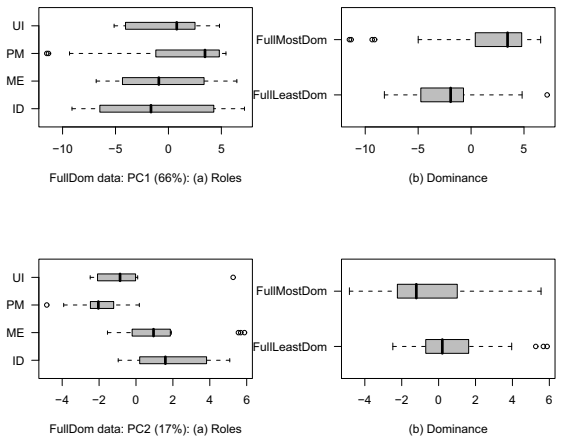

Figure 2: *Dominance in IDIAP meetings, first two principal components for Full most/least dominance agreement.*

Table 4: *IDIAP reduced data set: proportion of variance explained by the first two PCs. for Full most/least agreement.*

	% Var.	loadings
PC1	66.4	Energy, voicing_rate, −Hamm_effort, F_p
PC2	17.4	RCG, SKG, GOG, OQG

are in agreement with the perceptions reported in [7] and Table 2 about loud, tense voice for high dominant persons and soft, fearful voice for low dominance persons. Contradictory values for this perception were found for Hamm_effort and and OQG.

In [3] it has been predicted that for the reduced IDIAP corpus the PM is the most dominant person with an accuracy of 75% and 65% at human perception. This coincides with our findings that in these meetings PM speaks with a loud vocal effort, thus it can be concluded that PM performs as dominant person; similarly it can be argued that ME performs as less dominant person because he/she was found to speak with a more soft vocal effort. At human perception ME was found (majority agreement) the least dominant person in aprox. 43% of the cases.

6. Conclusions

In this paper we investigate the prosody and voice quality of dominance in scenario meetings. We have found that in these scenarios the most dominant person tends to speak with a louder-than-average voice quality and the least dominant person with a softer-than-average voice quality. We also found that the most dominant role in the meetings is the project manager and the least dominant the marketing expert.

In view of expressive speech synthesis, the identified acoustic measures could be used to specify high-level control over some social signals. The synthesized speech can thereby be requested to sound e.g. more dominant; rules informed by data such as the results of this and further work will transform such specifications into lower-level features governing acoustic correlates in the output (irrespective of whether waveform concatenation or vocoding techniques are employed).

In future developments we will carry out an in-depth analysis of whether the observed patterns can be solely attributed to intra-speaker differences, or whether they are related to speaker-specific effects; in this respect gradient normalisation like the one applied to some of the measures in this study will be considered for other measures.

7. Acknowledgements

This work is supported by the DFG project PAVOQUE and the EU project SSPNet (FP7/2007-2013).

8. References

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

[2] H. Hung *et al.*, "Using audio and video features to classify the most dominant person in a group meeting," in *ACM Multimedia*, 2007, pp. 835–838.

[3] D. B. Jayagopi, S. Ba, J. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Proc. 10th Int. Conf. on Multimodal Interfaces*, Chania, Crete, Greece, 2008, pp. 45–52.

[4] R. Rienks and D. Heylen, "Dominance detection in meetings using easily obtainable features," in *Proc. Workshop on Machine Learning for Multimodal Interaction*, 2006, pp. 76–86.

[5] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1775–1787, 2009.

[6] N. Dunbar and J. Burgoon, "Perceptions of power and interactional dominance in interpersonal relationships," *J. Soc. Pers. Relat.*, vol. 22, no. 2, pp. 207–233, 2005.

[7] J. E. Driskell, B. Olmstead, and E. Salas, "Task cues, dominance cues, and influence in task groups," *J. Appl. Psych.*, vol. 78, no. 1, pp. 51–60, 1993.

[8] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proc. Workshop on Machine Learning for Multimodal Interaction*, 2005, pp. 28–39.

[9] M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters under realworld disturbances," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006.

[10] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice quality," *Acta Otolaryngologica*, no. 90, pp. 441–451, 1980.

[11] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, 2004.

[12] C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, and S. Planet, "Discriminating expressive speech styles by voice quality parameterization," in *Proc. 16th Int. Congr. of Phonetic Sciences*, Saarbrücken, Germany, 2007.

[13] M. Charfuelan and M. Schröder, "Investigating the prosody and voice quality of social signals in scenario meetings," *submited to: Speech Commun. Special Issue on Sensing Emotion and Affect – Facing Realism in Speech Processing*, 2010.

[14] C. T. Ishi, H. Ishiguro, and N. Hagita, "Evaluation of prosodic and voice quality features on automatic extraction of paralinguistic information," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Beijing, China, 2006.

[15] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

[16] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, no. 1-2, pp. 189–212, 2003.

[17] P. Brown and S. C. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.