Investigating the prosody and voice quality of social signals in scenario meetings

Marcela Charfuelan, Marc Schröder

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany and
Alt-Moabit 91c, D-10559, Berlin, Germany
{firstname.lastname}@dfki.de

Abstract. In this study we propose a methodology to investigate possible prosody and voice quality correlates of social signals, and test-run it on annotated naturalistic recordings of scenario meetings. The core method consists of computing a set of prosody and voice quality measures, followed by a Principal Components Analysis (PCA) and Support Vector Machine (SVM) classification to identify the core factors predicting the associated social signal or related annotation. We apply the methodology to controlled data and two types of annotations in the AMI meeting corpus that are relevant for social signalling: dialogue acts and speaker roles.

Keywords: Prosody, Voice quality, Vocal social signals, Acoustic measures, Acoustic correlates, Perceptual interpretation

1 Introduction

The new research area of Social Signal Processing (SSP) is aimed at automatic understanding of social interactions through analysis of nonverbal behaviour. Social signals include (dis)-agreement, empathy, hostility, politeness, and any other stances towards others, and can be expressed through verbal and nonverbal means in different modalities [17]. One of the modalities through which social signals are supposedly expressed is *vocal nonverbal behaviour* – not *what* is said, but *how* it is said. This includes prosodic features such as pitch, energy and rhythm, as well as voice qualities such as harsh, creaky, tense, etc.

Most of recently reported works, related to the detection and classification of social signals, use only prosodic cues and in some cases in combinations with other cues. For example in [10] nonverbal prosodic and visual cues are used for predicting dominance and role-based status in scenario meetings. In [4] prosodic features have been used in combination with lexical and structural features for automatic detection of agreement in multiparty conversations. Prosodic features have been reported to discriminate quite well among dialogue acts [3] and voice quality features to discriminate quite well among emotions [11, 12]. Furthermore in [9], both prosodic and voice quality features have been used to identify some groups of speech acts expressing specific functions, emotion or attitude.

This paper addresses the question whether we can observe, in corpora of spontaneous interactions, any systematic effects of social signals on measures of

prosody and voice quality. If we are able to find such effects, we would like to know if prosody and voice quality carry redundant or complementary information, and whether the effects are perceptually interpretable. Our main goal is to develop a methodology for addressing these research questions, and to test-run it on a number of existing data sets. We start with controlled data which allow us to verify that the measures yield interpretations that are consistent with prior knowledge. We then proceed to apply the methodology to part of the Augmented Multi-party Interaction (AMI) meeting corpus, in which we investigate two types of existing annotations: dialogue acts and speaker roles. The paper is organised as follows. After outline the proposed methodology in Section 2, we describe in Section 3, the prosody and voice quality measures used in this study, as well as their perceptual interpretation. In Section 4 the methodology is test run on controlled data: the NECA database of voice quality and in Section 5, we present the application of the methodology to dialogue acts and roles in the AMI meeting corpus. Finally in Section 6 we draw conclusions and outline future work.

2 Outline of the methodology

The starting point for the analysis is a collection of speech recordings with associated annotations, afterwards the following steps are performed. **(1) Acoustic measures extraction** and if necessary, perform a pre-processing of the data. With annotations of spontaneous data it might be necessary to reduce the variability of the data, we exemplify two possible pre-processing methods in the work with dialogue acts and roles in Section 5. **(2) Analysis of Variance**, we compute a simple Analysis of Variance (ANOVA) for each of the features. This is a simple first assessment of whether we find significant effects among the annotations under analysis. **(3) Principal Component Analysis (PCA)**, performed on the acoustic features. PCA is used as a technique to reduce redundancy among the acoustic measures and to identify salient effects. In order to find systematic differences, we look at the distribution of our annotated data along the PCs. Visually, this distribution can be shown as a scatter plot of observations on the first two PCs; numerically, we can give means and standard deviations for annotation classes on the different PCs. In order to relate this distribution to perceptual interpretations, we can attempt to interpret the "meaning" of each PC in terms of the acoustic features with high loadings on a given PC. **(4) Classification**, in order to assess the quantitative distinctiveness of the acoustic features, we train a classifier to predict the annotations from the acoustic features. We chose Support Vector Machine (SVM) as a classifier. We train separate classifiers to predict the annotations from prosody features alone, from voice quality features alone, and from all features. Comparing these numbers allows us to determine if the acoustic measures are complementary or redundant.

3 Acoustic measures and their perceptual correlates

Table 1 shows the prosody and voice quality (VQ) measures used in this study. Prosody measures have been extracted frame-based and averaged per utter-

ance. VQ measures are extracted frame and utterance based. Frame-based VQ measures are rough spectral estimates of traditional voice quality parameters normally calculated in time domain. These measures were developed in [11] and tested successfully on classification of emotions under different levels of noise and reverberation. These measures are gradients (kind of normalisation by F0) instead of amplitud ratios and are calculated on the basis of frame-based raw measures like **formant frequencies:** F_1, F_2, F_3, F_4; **formant bandwidths:** B_1, B_2, B_3, B_4; **amplitude of the first two harmonics** at F0 and $2F0$: H_1, H_2; **frequency of spectrum peaks near formants**: F_{1p}, F_{2p}, F_{3p}; and **amplitude of spectrum peaks near formants**: A_{1p}, A_{2p}, A_{3p}. A tilde on some of the raw measures indicates that these measures have additionally included vocal tract influence compensation [11]. Utterance-based VQ measures were originally developed in [7] 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 long term average spectrum (LTAS) in three bands of frequency: 0-2kHz, 2-5kHz and 5-8kHz.

Type	Acoustic measure	Definition
Prosody	averageF0	Average fundamental frequency
	maxF0	Maximum F0
	minF0	Minimum F0
	rangeF0	maxF0 - minF0
	energy	Short term energy $\sum x^2$
	voicing rate	Number of voiced frames per time unit
VQ	**Frame-based [11]:**	
	OQG: Open Quotient Gradient	$(\tilde{H}_1 - \tilde{H}_2)/F0$
	GOG: Glottal Opening Gradient	$(\tilde{H}_1 - \tilde{A}_{1p})/(F_{1p} - F0)$
	SKG: Skewness Gradient	$(\tilde{H}_1 - \tilde{A}_{2p})/(F_{2p} - F0)$
	RCG: Rate of Closure Gradient	$(\tilde{H}_1 - \tilde{A}_{3p})/(F_{3p} - F0)$
	IC: Incompleteness of Closure	B_1/F_1
	Utterance-based [14, 7]:	
	Hamm_effort	LTAS_{2-5k}
	Hamm_breathy	$(\text{LTAS}_{0-2k}$ - $\text{LTAS}_{2-5k})$ - $(\text{LTAS}_{2-5k}$ - $\text{LTAS}_{5-8k})$
	Hamm_head	$(\text{LTAS}_{0-2k}$ - $\text{LTAS}_{5-8k})$
	Hamm_coarse	$(\text{LTAS}_{0-2k}$ - $\text{LTAS}_{2-5k})$
	Hamm_unstable	$(\text{LTAS}_{2-5k}$ - $\text{LTAS}_{5-8k})$
	slope_ltas1kHz	Least squared line fit of LTAS above 1 kHz in the log-frequency domain (dB/oct).

Table 1: Prosody and voice quality (VQ) measures used in this study.

3.1 Perceptual interpretation of acoustic measures

On the literature it is more common to find perceptual interpretations of traditional time domain voice quality measures. In the following we review the spectral effect and perceptual interpretation of traditional time domain voice quality measures and deduce the expected behaviour of their spectral domain counterparts.

Open quotient indicates the time during which the glottis is open and it is defined in the time domain as a fraction of the total glottal period. According to

[16] the primary acoustic manifestation of a narrow glottal pulse, i.e. of a decrease in open time, is a reduction of the amplitude of the fundamental component in the source spectrum relative to adjacent harmonics . Thus the spectral effect of the open quotient can be determined by the difference $(\tilde{H}_1 - \tilde{H}_2)$ [16, 6]. This means that a decrease in time domain open quotient corresponds also to a decrease in the spectral OQG. On the perceptual side, a very dominant H_1 has been widely found to be highly correlated with a breathy mode of phonation whereas a relatively strong H_2 can be correlated with tense or creaky voice [6].

Glottal opening and incompleteness of closure, glottal opening corresponds to the degree of opening over the entire glottal cycle. According to [16] the spectral effect of the glottal opening can be determined by the difference between the first harmonic and the amplitude of the first formant: $(\tilde{H}_1 - \tilde{A}_{1p})$. An increase in glottal opening correspond to a decrease in the amplitude of the first harmonic A_1 (increase in B_1) and therefore an increase of the spectral GOG and IC. On the perceptual side breathy voices have been associated with wide B_1 and tense voices with narrow B_1 [6], this means large values of GOG and IC for breathy voices and small values for tense voices.

Skewness and Rate of closure, skewness describes the abruptness (slope) of the glottal closure and the rate of closure the rate of decrease of flow at the instant of closure. [16] and [8] proposed that the amplitude of the third formant relative to that of the first harmonic $(H_1 - A_3)$ is a reasonably accurate indication of source spectral tilt, except if H_1 is weak. So for an increase in skewness and rate of closure we expect a decrease on the spectral SKG and RCG measures since the amplitude at middle and high frequencies increases relative to the amplitude at low frequencies. On the perceptual side, tense and creaky voices have been associated with high skewness or high speed quotient (SQ) and high rate of closure; on contrary breathy voices have been associated with low SQ [5].

Hamm_effort relates to vocal effort in a broad sense, *Hamm_breathy* will be expected larger for breathy voices than for creaky voices. *Hamm_head* is associated with head or chest register [7]. In [14] strong correlations of *Hamm_effort* have been found for the activation dimension of emotion. On the evaluation dimension, male speakers showed clear negative effects on the *Hamm_breathy* and *Hamm_coarse* for positive evaluation. On the power dimension, a negative correlation of power with *Hamm_unstable* has been found, i.e. higher power corresponds to a lower and more stable voice. Also in this study it has been found that higher activation corresponds to higher F0 median and range, larger F0 excursions and a flatter spectral slope, i.e, more high-frequency energy.

4 Test-running the methodology on controlled data: NECA vocal effort database

We analyse three levels of vocal effort in the NECA database [15], having in mind that vocal effort describes a broad range of voice qualities [13]. With this relatively simple database we aim, not only to verify that the acoustic measures have the expected discriminatory power, but also to assert and exemplify the

steps of the methodology proposed in this work. The NECA vocal effort database contains a full German diphone set for each of three levels of vocal effort ("soft", "modal" and "loud") and two speakers (one male, one female). Perception of the intended vocal effort was verified using stimuli generated using diphone synthesis voices built from the recordings [15]. Out of the original recordings, 100 short words per speaker and per vocal effort were used in this experiment (600 in total).

Analysis of Variance: the prosody and voice quality measures presented in Table 1 are extracted from the NECA database. After applying ANOVA on these measures it was found that most of the measures are significantly different among the three classes loud, modal and soft, at level 0.1% ($p<0.001$) except for: maxF0 ($p<0.01$) and averageF0 and minF0 ($p<0.3$).

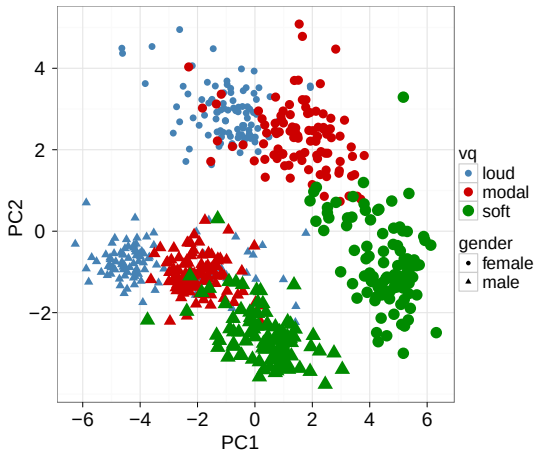


Fig. 1: NECA vocal effort PCA: variance explained PC1=45% and PC2=23%.

Principal Component Analysis: Figure 1 shows the three vocal effort clusters, loud, modal and soft for male and female speakers, obtained after PCA of both prosody and voice quality acoustic measures. It can be seen that both the speaker gender and the intended vocal effort are very well separated on this two-dimensional solution. The first three more loaded measures in PC1 are: Hamm_breathy, Hamm_coarse and RCG; and in PC2 are: maxF0, averageF0 and Hamm_head. Together these two PCs explain 68% of the variance. PC1 represents voice quality measures. PC2 is related to the fundamental frequency and the relative amount of high-frequency energy in the spectrum. Taking into account the perceptual interpretation of voice quality measures presented in Section 3.1 and the acoustic measures extracted from the NECA database, it was verified that the tendencies of the following measures, in terms of soft and loud levels relative to modal vocal effort, are consistent:

Soft vocal effort, (including here perceptions like breathy, whisper, lax voices) high OQG, high GOG, high SKG, high RCG, high IC, low Hamm_effort, high Hamm_breathy, low F0, flatter slope_ltas1kHz.

Loud vocal effort, (including here perceptions like tense, creaky, fry, pressed voice) low OQG, low GOG, low SKG, low RCG, low IC, high Hamm_effort, low Hamm_breathy, high F0, steeper slope_ltas1kHz.

Classification: three SVM models were trained with 60% of the NECA data using prosody and voice quality measures together and separately; 40% of the NECA data was used for testing. Table 2 shows the SVM average classification results for the annotation sets of the different data sets used in this study. The classification results for the three vocal effort levels in the NECA DB show that the voice quality measures produce a very good classification rate (90.8%) almost as good as using both prosody and voice quality measures (91.2%), which is the best classification rate for this database. The classification results obtained using just prosodic features is quite low in comparison to the others, maybe explained by the lack of significance on some of the prosodic measures.

Measures	Vocal effort NECA DB	Dialogue act AMI-IDIAP meetings	Role AMI-IDIAP meetings
Prosody + Voice quality	**91.2**	**44.3**	41.8
Prosody	77.5	42.8	19.0
Voice quality	90.8	42.8	**45.8**
Chance level	33.3	25.0	25.0

Table 2: SVM classification rate for the data sets used in this study using both prosody and voice quality measures and the two type of measures separately.

5 Analysis of meeting data: AMI Meeting corpus

The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings. Some of the meetings it contains are naturally occurring, and some are elicited, particularly using a scenario in which the participants play different roles. In this work elicited meetings are studied. In the scenario four participants play the roles: Project Manager (PM), Marketing Expert (ME), User Interface designer (UI) and Industrial Designer (ID) [1]. Nine meetings held at IDIAP Research institute (IS1000-IS1009, excluding IS1002) were selected from the AMI corpus, corresponding to 36 speakers (26 male and 10 female). The audio was taken from the individual headset. Table 3 presents the total number of dialogue act utterances extracted from these meetings and their distribution according to the dialogue act types studied in this work and speaker roles.

DA vs. Role	PM	UI	ID	ME	Total
Assess	704	605	632	768	2709
Elicit	489	215	187	362	1253
Suggest	509	442	412	396	1759
Inform	1343	1538	1314	1470	5665
Total	3045	2800	2545	2996	11386

Table 3: Distribution of dialogue act utterances extracted from the meetings.

DA vs. Role	PM	UI	ID	ME	Total
Assess	10	11	2	8	31
Elicit	15	6	5	10	36
Suggest	22	18	10	18	68
Inform	55	58	33	92	238
Total	102	93	50	128	373

Table 4: Distribution of dialogue act utterances containing the word “control”

5.1 Dialogue acts analysis

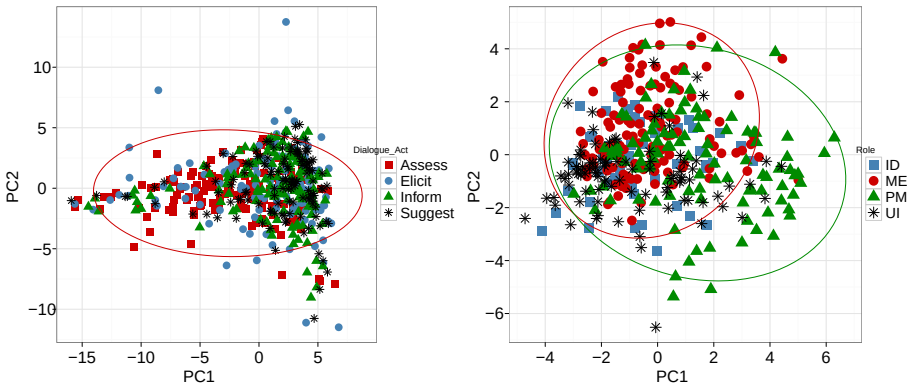
Features extraction and pre-processing: we have selected four frequently annotated dialogue acts that seem to have a clearly different meaning: Inform, Suggest, Assess and Elicit (grouping different elicit types). Our objective is to analyse variation or patterns on the measures due to dialogue acts or roles, but the measures are also affected by other sources of variation like speaker gender, individual speaking style, various sources of noise including overlapping speech, outbursts such as laughter, as well as the intrinsic contextual variability. When applying PCA directly on the measures per dialogue act or per role, as we did for the controlled data in Section 4, 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 dialogue act and speaker. This approach is comparable to the use of the Long-Term Average Spectrum (LTAS) as a means of "averaging out" the local effects of phonetic identity on the spectral distribution [7]. First of all the frame-based measures extracted from each dialogue act are averaged, resulting in one averaged frame-based measure per dialogue act. Then, all the dialogue act measures corresponding to a particular speaker in each sub-meeting are averaged.

Analysis of Variance: the analysis of variance corresponding to the averaged acoustic measures extracted from the four types of dialogue acts showed that the prosody measures and most of the voice quality measures allows to reject the null hypothesis that there is no significant difference among the dialogue act types. It seems that for this data the most relevant acoustic measures are the prosodic ones, significance level of 0.1 and 1%, although energy has a significance level of 5%. Three measures on the voice quality measures set: SKG, RCG and Hamm_unstable seem to be not significantly different.

Principal Component Analysis: Figure 2 (a) shows the projection of the averaged data onto a PC1-PC2 plane. It can be observed (indicated by an ellipse) that the first PC discriminates the Assess dialogue act from the others. The first three more loaded measures in PC1 are: Hamm_effort, Hamm_coarse and voicing rate; in PC2 are: RCG, GOG and SKG; and in PC3 are: averageF0, minF0 and maxF0. Analysing the mean acoustic values of these measures and referring to the perceptual correlates of Section 3.1, we observe that Hamm_effort for Assess has the highest value which indicates that Assess vocal effort is louder than the other DAs. The same loud vocal effort effect is observed for Hamm_coarse, and OQG, which have the lowest values among the DAs. AverageF0 and voicing rate present relatively small values which also suggest low activation. F0 was verified for both female and male data isolated, in both cases the lower effect for Assess DA was observed. The mean GOG value for Assess is not the expected for loud vocal effort, its value is higher than for the other DAs.

Classification: the classification of four types of dialogue acts in the AMI-IDIAP averaged data was performed speaker independent, with "leaving-one-speaker-out" cross validation and SVM. The classification results for dialogue

acts are presented in Table 2. When both prosody and voice quality measures are used the average classification rate is 44.3%, no improvement was observed when using prosody or voice quality features separately obtaining in both cases an average classification rate of 42.8%. Among the DAs, Assess appears as best classified (76.4%) confirming the salient tendency observed in the PCA analysis. A relatively good classification was obtained as well for Inform (56.6%) although this DA did not present a salient tendency in the PCA analysis.



(a) Dialogue acts in meetings (b) Roles in meetings

Fig. 2: AMI-IDIAP PCA: (a) Average data, variance explained PC1=49% and PC2=20%. (b) Single-word data, variance explained PC1=25% and PC2=18%.

5.2 Speaker roles analysis

Features extraction and pre-processing: in an attempt to reduce the variability of the data, with respect to roles, averaging per speaker and role was applied but just weak effects were observed. So in a second experiment intended to reduce variability we control the phonetic content. That is, we used only the different occurrences of a single frequent word, "control". It would have been preferable to investigate a single vowel, but time-aligned phonetic labels are not yet available in the AMI corpus. Table 4 shows the distribution of dialogue acts that contain the word "control" per dialogue act type and speaker role.

Analysis of Variance: the analysis of variance showed that the null hypothesis can be rejected because most of the prosody and voice quality measures among role types are significantly different at level 0.1%. The prosody measures minF0 and voicing rate have a significance level of 5% and rangeF0 of 1%. The voice quality measure Hamm_head has a significance level of 1%, so it seems that the voice quality features are the most relevant acoustic features in this data set.

Principal Component Analysis: Figure 2 (b) shows the projection of the single-word data onto a PC1-PC2 plane. Clusters for ME and PM roles are apparent that differ (indicated by ellipses) from the general distribution. Gender distribution of ME and PM shows that these roles 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 2 (b) that PM spreads more than average across PC2, whereas ME spreads more than average across PC1. This effect can not be explained merely by speaker gender, but seems specific for the speaker roles. The first three more loded features in PC1 are: Hamm_unstable, slope_ltas1kz and averageF0; in PC2: RCG, SKG and GOG; in PC3 Hamm_effort, Hamm_coarse and SKG. Analysing the mean acoustic values of these measures we observe that the main salient indicators of PM when comparing to the other roles are: a higher value of averageF0 and a lower value of GOG which suggest a loud vocal effort tendency employed by PM. Salient indicators of ME when comparing to the other roles are: higher values for GOG, SKG and RCG which suggest a soft vocal effort tendency employed by ME. The mean value of Hamm_unstable for ME seems to be in a modal range. The spectral slope above 1 kHz value for ME is relatively flat when comparing to 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.

Classification: the classification of four roles in the AMI-IDIAP single-word data was also performed speaker independent, with "leaving-one-speaker-out" cross validation and SVM. The classification results for roles are presented in Table 2. In this case the best classification result is obtained when using just voice quality features (45.8%). When both prosody and voice quality measures are used the average classification rate is 41.8% and the classification rate drops to 19.0% when using just prosody measures. Among roles the best classification rates are for ME (61.72%) and PM (53.92%). The salient tendencies of ME and PM are confirmed with these results.

6 Conclusions

In this paper we have presented a methodology for investigating prosody and voice quality features of social signals in naturalistic recordings. We combine simple prosodic measures with measures from the literature that were reported to capture voice quality robustly, and use Principal Components Analysis and Support vector Machine classification to factor out redundancy and identify the strongest effects. The robustness of the measures and the methodology employed in this study have been verified with controlled data. We have verified that the results are consistent with perceptual impressions. The systematic differences found in the three sets of data are mainly concern with both prosody and voice quality measures. Using the methodology proposed in this paper it has been found that: "assess" dialogue acts were often spoken with a louder vocal effort than other dialogue acts; marketing experts often spoke with a softer voice than average, whereas project managers often spoke with a louder voice than average. So we would expect that when new databases with annotations for specific types of social signals become available in the future, the methodology can also be applied to these data sets, as it was the case for the analysis of dominance in some AMI-IDIAP meetings [2].

Future work will extend the set of acoustic measures to include other acoustic measures such as contour shapes and spectral measures. Another line of work

will be to extend the methodology to be able not only to detect general effects on the data but more local effects, or salient acoustic events in a meeting.

Acknowledgements. The research leading to these results has received funding from the EU Programme FP7/2007-2013, under grant agreement no. 231287 (SSPNet).

References

1. Carletta, J., et al.: The ami meeting corpus: A pre-announcement. In: Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI'05). pp. 28–39. LNCS Springer-Verlag (2005)
2. Charfuelan, M., Schröder, M., Steiner, I.: Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings. In: Proc. Interspeech. Makuhari, Japan (2010)
3. Fernandez, R., Picard, R.W.: Dialog act classification from prosodic features using support vector machines. In: Proc. Speech Prosody. Aix-en-Provence, France (2002)
4. Germesin, S., Wilson, T.: Agreement detection in multiparty conversation. In: Proc. ICMI-MLMI'09. Cambridge, Massachusetts, USA (2009)
5. Gobl, C., Chasaide, A.N.: The role of voice quality in communicating emotion, mood and attitude. Speech Commun. 40(1-2), 189–212 (2003)
6. Gobl, C., Chasaide, A.N.: Voice source variation and its communicative functions, pp. 378–423. The Handbook of Phonetic Sciences (second edition) (2010)
7. Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., Wedin, L.: Perceptual and acoustic correlates of abnormal voice quality. Acta Otolaryngologica (90) (1980)
8. Hanson, H.M.: Glottal characteristics of female speakers: Acoustic correlates. The Journal of the Acoustical Society of America 101(1), 466–481 (1997)
9. Ishi, C.T., Ishiguro, H., Hagita, N.: Evaluation of prosodic and voice quality features on automatic extraction of paralinguistic information. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. Beijin, China (2006)
10. Jayagopi, D.B., Ba, S., Odobez, J., Gatica-Perez, D.: Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In: Proc. 10th ICMI'08. pp. 45–52. Chania, Crete, Greece (2008)
11. Lugger, M., Yang, B., Wokurek, W.: Robust estimation of voice quality parameters under realworld disturbances. In: IEEE ICASSP. Toulouse, France (2006)
12. Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X., Planet, S.: Discriminating expressive speech styles by voice quality parameterization. In: Proc. 16th Internat. Cong. of Phonetic Sciences (ICPhS). Saarbrücken, Germany (2007)
13. Nordstrom, K., Tzanetakis, G., Driessen, P.: Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction. IEEE Transactions on Audio, Speech and Language proscessing 16(6) (2008)
14. Schröder, M.: Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. Ph.D. thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University (2004)
15. Schröder, M., Grice, M.: Expressing vocal effort in concatenative synthesis. In: Proc. 15th Internat. Cong. of Phonetic Sciences (ICPhS). Barcelona, Spain (2003)
16. Stevens, K., Hanson, H.: Classification of glottal vibration from acoustic measurements, chap. 9. No. 147-170, Vocal Fold Physiology: Voice Quality Control (1994)
17. Vinciarelli, A., Salamin, H., Pantic, M.: Social signal processing: Understanding social interactions through nonverbal behavior analysis. In: IEEE Computer Vision and Pattern Recognition Workshops. pp. 42–49 (2009)