CLASSIFICATION OF LISTENER LINGUISTIC VOCALISATIONS IN INTERACTIVE MEETINGS

Marcela Charfuelan, Marc Schröder, Sathish Pammi

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany and
Alt-Moabit 91c, D-10559, Berlin, Germany
{marcela.charfuelan, marc.schroeder, sathish.pammi}@dfki.de

ABSTRACT

This paper presents the classification of two types of listener linguistic vocalisations that occur during spontaneous interactions in the AMI-IDIAP meeting corpus. In a first stage, principal component analysis (PCA) of low level acoustic measures is used to separate salient lower and higher acoustic events. We have found that two types of linguistic vocalisations appear very often in salient events. Among the lower salient acoustic events 44% correspond to backchannel vocalisations whereas among the higher salient events 32% correspond to stall vocalisations. In a second stage, once salient acoustic events are split into high and low two Support Vector Machine (SVM) classifiers are trained with different acoustic features to classify these two sets separately. We have got a classification accuracy of 81% and 80% for stall and backchannel linguistic vocalisations. The approach can be applied on the development of SAL (sensitive artificial listener) systems or interative systems in general.

1. INTRODUCTION

Automatic understanding and synthesis of nonverbal communication is becoming increasingly important in computing technologies like human computer interaction (HCI), embodied conversational agents (ECAs), Social signal processing etc. [18]. Face-to-Face meetings or small group conversations have been extensively used for studying this type of communication, i.e. communication that includes nonverbal behavioural cues like facial expressions, vocalisations, gestures, postures, etc. [6]. Vocal nonverbal behaviour cues accounts for *how* something is said not *what* is said and according to Richmond et al. includes five major components: prosody, linguistic and non-linguistic vocalisations, silences and turn-taking patterns [15]. Examples of typical linguistic vocalisations are: "yeah", "uh-huh", "hm", to signal that the listener hears what the speaker is talking (backchannel) and "uh", "um", "so", to indicate that the speaker is about to speak or wants to keep or cede the floor (stalling). Non-linguistic vocalisations includes cries, laughs, shouts, yawns, sobbing etc. typically related to strong emotional states or tight social bounds [18].

In this work we analyse the prosody, voice quality and spectral characteristics of two types of annotated non-intentional dialogue acts in the AMI-IDIAP meeting corpus: backchannel and stall. Backchannel utterances are expected to be not so acoustically salient because most of the time this kind of utterances are not intended to interrupt but simply signal to the speaker listening or attention [19]. On the other hand nonverbal vocalisations intended to grab the floor or indicate that the speaker is about to speak might be more acoustically salient. In fact Clark et al. [5] has reported that speakers formulate "uh" and "um" with a prosody that makes them distinguishable from the surrounding word when placed within intonation units [5]. Therefore we have decided to concentrate the analysis on salient lower and higher acoustic events. In this work, salient lower and higher acoustic events are determined by a principal component analysis (PCA) of low level acoustic measures extracted from voiced frames. We have found that in fact the acoustic and perceptual characteristics of backchannel and stall vocalisations are quite different and not only correlated with prosody but also with voice quality and spectral characteristics. These two types of vocalisations appear as the more frequent in salient acoustic events, 44% of the lower salient events correspond to backchannel utterances and 32% of the higher salient events correspond to stall utterances. Thus once salient acoustic events are split different sets of acoustic features are used to train two Support Vector Machine (SVM) classifiers, one for higher salient events and another for lower salient events. We have got a classification accuracy of 81% and 80% for stall and backchannel linguistic vocalisations.

Several works have reported on the detection of cues and context of backchannel responses specially to train models for predicting when an interactive system should generate backchannel [17, 7]. There exist few works on the detection and classification of different types of listener linguistic vocalisations, related work is [2], where the prosody of backchannels in American English is analysed. The classification of this type of vocalisations would be of special interest to improve the response of interactive systems, so for example, an interactive system can simply acknowledge and continue talking when detecting backchannel and perhaps politely give the floor when detecting stall.

In Sections 2 and 3 the corpus and acoustic measures used in this work are described. The detection of salient acoustic events and their distribution among dialogue acts are explained in Section 4. Classification results for salient acoustic events and the types of acoustic measures that better discriminate lower and higher salient events are presented in Section 5. Conclusions and future work are presented in Section 6.

2. CORPUS AND METHOD

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 of employees in an electronics company that decides to develop a new type of television remote control. Each meeting is organised in four phases (sub-meetings): (a) project kick-off, (b) functional design, (c) conceptual design and (d) detailed design, where the same group of four people participate [3].

This corpus contains recordings of both video and audio data, transcriptions and several levels of annotations, for example dialogue acts. The transcriptions include word level segmentation time-aligned to the recordings. The dialogue act (DA) annotations in the AMI corpus code speaker intentions according to: acts about the *information exchange* (inform, elicit-inform), acts about possible *actions* (suggest, offer, elicit-offer-or-suggestion), *comments* on previous discussion (assess, comment-about-understanding, elicit-assessment, elicit-comment-about-understanding), *social acts* (be-positive, be-negative) and the classes *other and fragment* which are bucket classes where the speaker is conveying an intention that do

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

not fit into any of the other classes. There is also a special class of dialogue acts that are actually non-intentional acts: backchannel and stall. Although these last two types of dialogue acts in the AMI corpus do not convey a precise intention they are very important and frequently used in spoken interaction. For example backchannel signals understanding (whether the listener understands the utterance of the speaker), attentiveness (whether the listener is attentive to the speech of the speaker), attitude, affect etc. [14]. Some examples of backchanel utterances in the AMI corpus are: "uhhuh", "mm-hmm", "yeah", "yep", "ok", "ah", "huh", "hmm", "mm". Examples of stall are "uh", "um", "so", these dialogue acts in the AMI corpus are special sounds called "filled pauses", used by people when they start speaking before they are really ready, or keep speaking when they have not figured out what to say, just to try to get or keep the attention of the group [1].

2.1 Method

Nine meetings held at the 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. From these meetings just the sections where the four participants interact or discuss actively are selected. Acoustic analysis is performed for the utterances corresponding to each annotated dialogue act in these sections. All intentional dialogue acts, except social dialogue acts (for which there are very few examples), are grouped and considered in this study as single classes. Among the non-intentional dialogue acts backchannel and stall are selected, other and fragment are not considered because they do not convey a clear intention. Table 1 presents the total number of dialogue act utterances extracted from the meetings and actually used in the analysis of salient events.

Intention	Dialogue Act	Extr.	Used
Inform	Inform	2541	
	Elicit-Inform	391	2932
Action	Suggest	850	
	Offer	93	
	Elicit-Offer-Or-Suggestion	69	1012
Comment	Assess	1511	
	Comment-About-Underst.	214	
	Elicit-Assessment	157	
	Elicit-Comment-Underst.	32	1914
Social	Be-Positive	173	
	Be-Negative	4	-
	Backchannel	1506	1506
	Stall	669	669
	Fragment	809	-
	Other	121	-
Total		9140	8033

Table 1: Distribution of dialogue act utterances in interactive sections of nine AMI-IDIAP meetings. Backchannel comprises 16.4% of the total data and Stall comprises 7.3% of the total data.

For each utterance used in this work acoustic measures are extracted at frame and utterance level (as explained in section 3). It has been reported that speakers formulate "uh" and "um" with a prosody that makes them distinguishable from the surrounding word when placed within intonation units [5], also it has been shown that energy and f0 are good indicators of "hot spots" or regions in which participants are highly involved in the discussion [20]. Therefore in this work the first analysis intended to detect salient vocal acoustic moments is concentrated on voiced frames, which convey most of the energy in speech. Principal component analysis of several low level measures is performed in order to find features that contribute the most to the variance and determine a threshold for salient higher and lower events. Thus salient higher and lower acoustic events are defined as all the dialogue acts whose first principal component value goes above and below a threshold that depends on a factor κ, empirically designed, and the quartile statistics of PC1:

$$Event_{low} < (\kappa * Q_1) \leqslant Q_2 \leqslant (\kappa * Q_3) < Event_{high} \qquad (1)$$

where:

κ : factor empirically designed

Q_1 : first quartile of PC1

Q_2 : second quartile of PC1 = median(PC1)

Q_3 : third quartile of PC1

Salient lower and higher points selected in this way are then used to train two SVM classifiers with which to classify both sets of data according to backchannel, stall or other dialogue acts.

3. ACOUSTIC MEASURES

The following acoustic measures have been extracted from all dialogue act utterances. Low level acoustic measures are extracted at frame level, frame length 25 ms. and frame shift 5 ms. The frame based measures are averaged per utterance. Prosody and voice quality measures are extracted at utterance level.

3.1 Low level acoustic measures

- *Voicing strengths:* [4, 11] are estimated with peak normalised cross correlation of the input signal. The correlation coefficient for a delay t is defined by:

$$c_t = \frac{\sum_{n=0}^{N-1} s(n)s(n+1)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n+t)}} \qquad (2)$$

  In this study one full band and five bandpass voicing strengths are calculated, that is, the input signal is filtered into five frequency bands, with pass-bands 0-1kHz, 1kHz-2kHz, 2kHz-4kHz, 4kHz-6kHz and 6kHz-8kHz and voicing strengths are calculated for each filtered signal.
- *Pitch harmonics magnitude:* [4, 11] corresponds to the Fourier magnitude of the first ten pitch harmonics of the residual signal obtained by inverse filtering. The Fourier magnitudes capture the shape of the excitation pulse, so it is expected that they capture variations in phonation types.
- *Spectral features*:
  - Mel-cepstral coefficients [9].
  - Spectral entropy, this is a kind of "peakiness" of the spectrum. This value is calculated as follows [12]: the spectrum X is converted into a Probability mass function (PMF) normalising it by:

$$x_i = \frac{X_i}{\sum_{i=1}^{N} X_i} \qquad i = 1 \quad to \quad N \qquad (3)$$

    where X_i is the energy of the i^{th} frequency component of the spectrum, x is the PMF of the spectrum and N is the number of points in the spectrum. Entropy for each frame is calculated by:

$$H(x) = -\sum_{x \in X} x_i * log_2 x_i \qquad (4)$$

    It is expected that spectral entropy is higher for voiced frames than for unvoiced, therefore it has been used in speech endpoint detection and in classification of emotions. In this work spectral entropy is calculated for the full band signal and for the five frequency bands filtered signals.

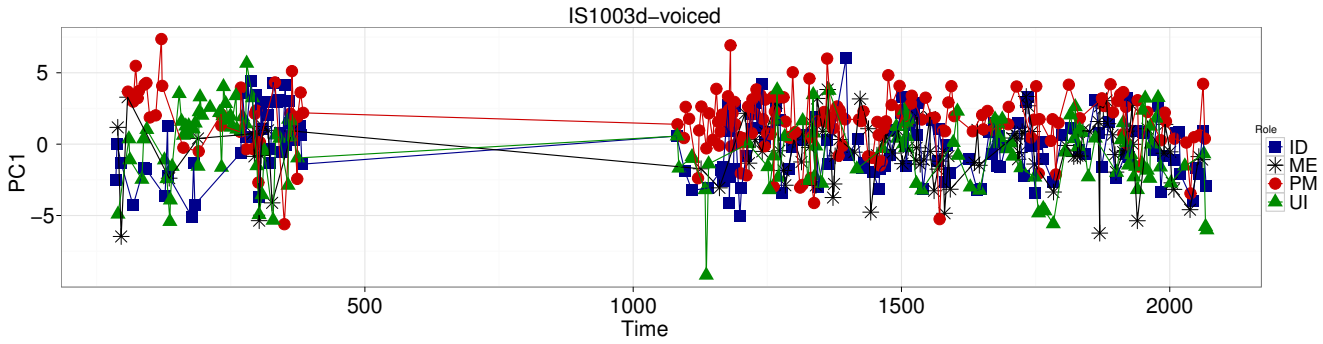


Figure 1: First principal component (PC1) Vs Time for voiced frames of meeting IS1003, phase d. The roles of the participants are: Project Manager (PM), Marketing Expert (ME), User Interface designer (UI) and Industrial Designer (ID)

- *Articulatory-based features*
  - Formants: F_1, F_2, F_3, F_4
  - Formant bandwidths: B_1, B_2, B_3, B_4
  - Formant dispersion: calculated as:

$$FD = \frac{(F2-F1)+(F3-F2)+(F4-F3)}{3} \quad (5)$$

3.2 Prosody acoustic measures

- Fundamental frequency or pitch (f0)
- Pitch entropy (calculated as the spectral entropy)
- maximum, minimum, and range of f0
- Duration of the utterance in seconds
- Voicing rate calculated as the number of voiced frames per time unit
- Energy, calculated as the short term energy $\sum x^2$

3.3 Voice quality acoustic measures

The following voice quality measures were originally developed by [8] and have been also used in emotion research by [16] and [13]. These measures are based on the calculation of the long term average spectrum (LTAS) in three bands of frequency: LTAS between 0-2kHz ($LTAS_{0-2k}$), LTAS between 2-5kHz ($LTAS_{2-5k}$) and LTAS between 5-8kHz ($LTAS_{5-8k}$). For each of these bands the maximum level is selected.

- 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 = ($LTAS_{2-5k}$ - $LTAS_{5-8k}$)

LTAS defined as:

$$LTAS_{dB}(f) = 10Log_{10}[\frac{1}{L}\sum_{i=1}^{L} PSD_i(f)] \quad (6)$$

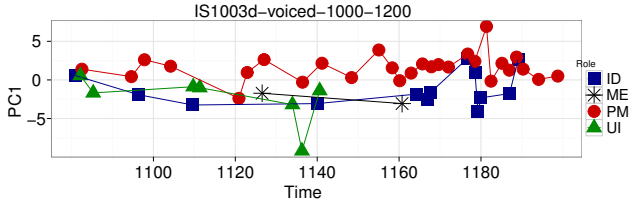
where $PSD_i(f)$ is the power spectral density of the i-th windowed frame of the signal.

Other voice quality measures used in this study are:

- slope_ltas: least squared line fit of LTAS in the log-frequency domain (dB/oct)
- slope_ltas1kz: least squared line fit of LTAS above 1 kHz in the log-frequency domain (dB/oct)
- slope_spectrum1kz: least squared line fit of spectrum above 1 kHz (dB/oct)

4. SALIENT ACOUSTIC EVENTS DETECTION

A first PCA was performed to find the low level acoustic features that contribute the most to the variance, this is determined by the strongest positive and negative loadings of the first principal component. Among the strongest positive loaded features are the *five sub-band voicing strengths* and among the strongest negative loaded are the *five sub-band spectral entropy features*. The 23 more loaded features were selected to perform the PCA with which the data is split, these features account for 29% of the variance on the first principal component.



Role	StartT	EndT	DialogueAct	PC1	Transcription
UI	1085.2	1089.6	Assess	-1.8	Okay.
UI	1109.7	1109.9	Backchannel	-1.6	Mm.
UI	1111.3	1111.5	Backchannel	-0.6	Okay.
UI	1133.9	1134.3	Backchannel	-4.0	Mm-hmm.
UI	**1136.2**	**1136.5**	**Backchannel**	**-9.6**	**Mm-hmm .**
PM	1136.4	1141.1	Suggest	-0.3	And if not if it's not the case y you would have ...
UI	1140.6	1140.7	Backchannel	-1.1	Yeah.
			...		
PM	1172.0	1175.4	Elicit-Inform	2.0	Do you think one would be enough, or such as
PM	1176.7	1178.4	Suggest	2.9	as number of branches?
PM	1178.4	1181.3	Elicit-Assess.	1.8	Three?
PM	**1181.3**	**1182.4**	**Stall**	**6.8**	**So,**
PM	1182.4	1183.1	Elicit-Off-Sugg.	-0.0	electronic.
PM	1185.1	1186.9	Elicit-Assess.	1.9	Single simple chip on print?

Figure 2: First principal component (PC1) points and corresponding transcriptions in segment 1000-1200 seconds of meeting IS1003d: the transcription corresponding to the lower UI point and the higher PM point are indicated in bold.

4.1 Temporal analysis

After performing a principal component analysis with the features that more contribute to the variance, the distribution of the first principal component points over time for the four participants in the meetings can be analysed. Figure 1 shows a plotting of the first principal component versus time for the phase "d" of the IS1003 meeting. As can be observed most of the points appear in the midle (approx. $-5 < PC1 < 5$), but there are also clear points that appear as salient lower ($PC1 < -5$) or higher ($PC1 > 5$). As an example, let us consider the transcriptions corresponding to the same phase "d" of the meeting IS1003 between the times 1000 and 1200 sec-

onds, Figure 2. The lower PC1 point with value -9.6 at time 1136.2 seconds corresponds to a backchannel utterance "Mm-hmm"; the higher PC1 point with value 6.8 at time 1181.3 corresponds to a stall utterance "So".

4.2 Salient events distribution

Figure 3 shows the percentage distribution of salient higher and lower events in the AMI-IDIAP meetings. It has been found empirically that with $\kappa = 2$ there exist a tendency of two types of dialogue acts to appear as the more frequent in salient acoustic events: 44% of the lower salient events correspond to backchannel utterances and 32% of the higher salient events correspond to stall utterances. A more detailed distribution of salient events for the other dialogue act groups is presented in Table 2.

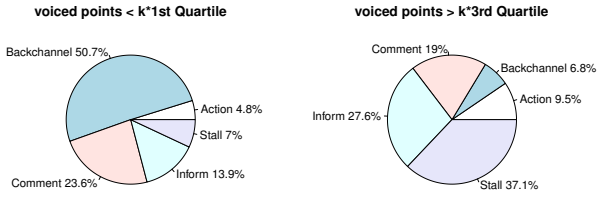


Figure 3: Distribution of lower and higher salient PC1 points corresponding to 17.8% of the total data, split with $\kappa = 2$.

Dialogue Act	Higher salient	Lower salient	Total
Inform	145	128	273
Action	50	44	94
Comment	100	217	317
Backchannel	36	465	501
Stall	195	64	259
Total	526	918	1444
% Total data	6.5	11.4	17.9

Table 2: Distribution of higher and lower salient acoustic vocal events in the AMI-IDIAP meetings, split with $\kappa = 2$.

4.3 Main acoustic features of salient events

PCA of the two sets of salient acoustic events was performed to determine which features discriminate better each set. The more loaded features for each set are presented in Table 3, including for comparison some prosody features (dur_seconds, f0 and energy) for which the loadings were very low. It is interesting to notice that prosody features are not so important, as discriminator features, for both sets, specially for lower salient events. Voice quality features in combination with spectral features seem to be the more discriminant features in both sets, specially for lower salient events. This might be expected since backchannel responses like "Mm", "Mm-hmm" are mostly unvoiced speech.

5. SALIENT ACOUSTIC EVENTS CLASSIFICATION

As it has been shown the acoustic characteristics of backchannel and stall dialogue acts in the AMI-IDIAP corpus are quite different at level of voiced frames. In this section classification of these two types of vocalisations is performed using prosody, voice quality and spectral features where the best combination of features is investigated. Three dialogue act categories are considered: backchannel, stall and other; here the other class includes all the intentional dialogue act groups except social (see Table 1). The classification of these three categories is performed in two stages. In a first classification stage the data is split into higher and lower acoustic events as

PC1 Higher salient		PC1 Lower salient	
Feature	loading	Feature	loading
slope_ltas1kz	0.29	slope_ltas	0.22
slope_spectrum1kz	0.28	mcep17	0.22
slope_ltas	0.28	mcep19	0.21
Hamm_breathy	0.24	mcep15	0.20
mcep2	0.20	mcep20	0.20
mcep7	0.19	mcep18	0.19
dur_seconds	0.12	dur_seconds	0.18
f0	0.10	energy	0.01
energy	0.07	f0	0.0002
voicing_rate	-0.16	voicing_rate	-0.22
mcep1	-0.23	Hamm_coarse	-0.23
Hamm_head	-0.26	mcep1	-0.25
Hamm_unstable	-0.28	Hamm_head	-0.27

Table 3: First principal component loadings for PCA of higher and lower salient events sets, split with $\kappa = 2$.

presented in section 4. In a second stage, higher and lower events are classified separately using two SVM classifiers trained with different feature sets. In the higher events classifier stall and other dialogue acts are classified and in the lower events classifier backchannel and other dialogue acts are classified.

This approach is motivated by the finding that the two types of vocalisations are the most frequently higher or lower acoustic events ($\kappa = 2$) on the studied meetings and also by the investigations of [10] where several stages of classification, with different features in each stage, have been shown to perform better in discriminating emotions. Of course the classification of the first stage introduce errors because not all the backchannel dialogue acts are in the lower salient events set and those that appear in the higher events set will be miss-classified. The same is the case for stall dialogue acts, but it is expected that in an application, for example, the miss-classification of these not so salient events in each class can be ignored as long as the more salient events are better detected and classified.

We have split the data as higher and lower acoustic events with different values of κ. The classification results presented in Table 4 corresponds to a $\kappa = 0$, that is, the median of PC1 is used to split the data. The table shows classification results for three SVM classifiers:ALL, HIGH and LOW trained with different feature sets. ALL classifies all the data as backchanel, stall or other; HIGH classifies the salient higher acoustic events as stall or other; LOW classifies the salient lower acoustic events as backchannel or other. In each case 20 trials of classification leaving-one-speaker-out cross validation with 200 randomly selected samples for each class are performed. The average classification after 20 trials is shown.

Features		% Avg. accuracy		
Type	No.	ALL	HIGH	LOW
prosody	9	63.9	76.3	75.7
vq	11	57.8	72.4	73.9
vq+prosody	20	64.9	78.7	77.6
vq+prosody+spectral	55	66.7	79.9	79.0
more discriminant features	23	61.0	79.0	70.4
PCA all features	14	66.3	81.2	76.7

Table 4: Classification accuracy of all dialogue acts and lower and higher salient events according to: Backchannel, Other and Stall. Data split with $\kappa = 0$ i.e. the median of PC1 is used to split the data.

The classification results show that the average classification of all the data without separation of salient events gives lower results (67.1%) than classifying the two sets separately (without taking into account the miss-classification introduced by the initial splitting of

	ALL			HIGH		LOW	
	Back.	Other	Stall	Other	Stall	Back.	Other
Back.	**72.6**	13.6	19.1	-	-	**80.7**	22.7
Other	9.7	**66.6**	19.9	**78.3**	18.6	19.3	**77.3**
Stall	17.5	19.7	**60.9**	21.7	**81.5**	-	-

Table 5: Confusion matrix for best classification results in Table 4 using vq+prosody+spectral features. Data split with $\kappa = 0$ i.e. the median of PC1 is used to split the data.

the data). The best discrimination for higher and lower events is obtained when prosody, voice quality and spectral features are used. In this case the classification accuracy of each class is: 81.5% for stall utterances and 80% for backchannel utterances (see more detailed results in Table 5). The more discriminant features used to split the data into two sets, give a good average classification result for HIGH but not so good for LOW. In this last case the addition of spectral features produce a better classification result. It has been observed that the best results can be obtained by using the first 14 principal components of a PCA performed with all the features. This makes a significant reduction on the number of features used. The classification results for the classifiers HIGH and LOW were repeated with $\kappa = 1$ and $\kappa = 2$, that is, with a reduced number of salient events. The classification results shown similar tendencies, so in an application the level of precision with which salient vocalisations are "attended" can be controlled by a particular value of κ.

6. CONCLUSIONS

In this study, we have presented a novel approach for detection and classification of two types of linguistic vocalisations that occur during spontaneous interactions in meetings. The approach is based on first detecting salient acoustic events using PCA of low level acoustic features extracted from voiced frames, and then classifying separately higher and lower acoustic events with two SVM classifiers trained with prosody, voice quality and spectral acoustic features. We have obtained an average classification of 80% for higher events and 79% for lower events.

The approach is motivated by the finding that the two types of vocalisations are the most frequently higher or lower acoustic events on the studied meetings and also by investigations that demonstrate that discrimination and classification of emotions (which can be considered salient acoustic events) is better performed when using several classification stages, being the first one binary classification according to two activation levels: high and low [10].

The first stage classification introduce errors because a few percentage of vocalisations will be miss-classified on the first splitting of the data, but it is expected that in an application the miss-classification of these not so salient events in each class can be ignored as long as the more salient events are better detected and classified. The initial splitting of the data has been tested using various thresholds $\kappa = 0, 1, 2$, giving similar results for the classification of salient higher and lower events. This property can be useful in an application where it would be interesting to control the level with which salient vocalisations are "attended", for example in a sensitive artificial listener (SAL) system. As future work, the generalisation of the approach will be tested in other meeting corpus and the possible co-occurrence of salient acoustic events with other nonverbal behaviour cues like gestures will be investigated.

REFERENCES

[1] AMI project consortium. Guidelines for dialogue act and addressee. annotation version 1.0. http://mmm.idiap.ch/private/ami/annotation/dialogue-_acts_manual_1.0.pdf, 2005.

[2] S. Benus, A. Gravano, and J. Hirschberg. The prosody of backchannels in american english. In *Proc. 16th International Congress of Phonetic Sciences (ICPhS)*, 2007.

[3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, K. J., V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. I., W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. In *Proc. Workshop on Machine Learning for Multimodal Interaction*, pages 28–39, 2005.

[4] W. C. Chu. *Mixed excitation linear prediction*, chapter 17, pages 454–485. Speech coding algorithms Foundations and Evolution of Standardized Coders. Wiley, 2003.

[5] H. Clark and J. Tree. Using uh and um in spontaneous speaking. *Cognition*, 84:73–111, 2002.

[6] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vis. Comput.*, 27(12):1775–1787, 2009.

[7] A. Gravano and J. Hirschberg. Backchannel-inviting cues in task-oriented dialogue. In *Proc. Interspeech*, Brighton, UK, 2009.

[8] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin. Perceptual and acoustic correlates of abnormal voice quality. *Acta Otolaryngologica*, (90):441–451, 1980.

[9] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H. Zen. Speech signal processing toolkit (SPTK), Version 3.4. http://sp-tk.sourceforge.net, 2009.

[10] M. Lugger and B. Yang. *Psychological Motivated Multi-Stage Emotion Classification Exploiting Voice Quality Features*, chapter 22, pages 395–410. I-Tech. Speech Recognition, Technologies and Applications, Vienna, Austria, 2008.

[11] V. McCree, A. *Low-Bit-Rate Speech Coding*, chapter 16, pages 337–341. Springer Handbook of Speech Processing. Springer, 2007.

[12] H. Misra, S. Ikbal, S. Sivadas, and H. Bourlard. Multi-resolution spectral entropy feature for robust ASR. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[13] 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 (ICPhS)*, Saarbrücken, Germany, 2007.

[14] D. Reidsma, K. P. Truong, H. van Welbergen, D. Neiberg, S. Pammi, and I. A. de Kok. Continuous interaction with a virtual human. In *Proc. of the 6th International Summer Workshop on Multimodal Interfaces (eNTERFACE10)*, pages 24–39, Amsterdam, 2010.

[15] V. Richmond and J. McCroskey. *Nonverbal behaviors in interpersonal relations*. Allyn & Bacon, 1995.

[16] M. Schröder. *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, 2004.

[17] K. P. Truong, R. Poppe, and D. Heylen. A rule-based backchannel prediction model using pitch and pause information. In *Proc. Interspeech*, Makuhari, Japan, 2010.

[18] A. Vinciarelli and G. Mohammadi. Towards a Technology of Nonverbal Communication: Vocal Behavior in Social and Affective Phenomena. In *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, pages 133–156. 2011.

[19] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32:1177–1207, 2000.

[20] B. Wrede and E. Shriberg. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proc. of Eurospeech*, Geneva, 2003.