Recognition of Interest in Human Conversational Speech

Björn Schuller, Niels Köhler, Ronald Müller, and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
{sch, mur, ri}@mmk.ei.tum.de

Abstract

Recognition of interest of a speaker within a human dialog bears great potential in many commercial applications. Within this work we therefore introduce an approach that analyses acoustic and linguistic cues of a spoken utterance. A systematic generation of more than 5k hi-level features basing on prosodic and spectral feature contours by means of descriptive statistical analysis and subsequent feature space optimization is used to find relevant acoustic attributes. For linguistic information integration a bag-of-words representation is used relying on a speech recognizer's output. One main aspect is the database of more than 2k spontaneous sub-speaker turns recorded and annotated for this analysis. Several influence factors as microphone distance and ASR versus annotation of spoken content are discussed. Overall remarkable performance of a running prototype can be reported discriminating between three levels of interest.

1. Introduction

Knowledge of a communication partner's interest possesses great potential in many commercial applications. Similar to the work introduced in [7] we are likewise interested in curiosity detection e.g. for topic switching, but based on speech analysis considering acoustic and linguistic cues. In order to quantify a speaker's interest we introduce three levels of interest (LOI) reaching from LOI=0 representing *disinterest*, *indifference*, and *neutrality* over LOI=1 standing for *light interest* to LOI=2 representing *strong interest*. The paper is structured as follows: after a short description of data collection we describe acoustic and linguistic feature computation, feature space optimization and classification, and finally provide experimental results and conclusions.

2. Database

In order to obtain a high number of spontaneous data in view of a speaker's interest, a data collection was carried out with 11 male subjects ranging from 23a to 45a, mean 29.7a, and 10 female subjects ranging from 20a to 57a, mean 30.1a. In the chosen scenario setup, an experimenter leads a subject through an interactive presentation of commercial products. The subject's role is to listen to explanations and topic presentations of the experimenter, ask several questions of his/her interest, and actively interact with the experimenter especially considering his/her interest to the addressed topics without respect to politeness. The conversation language is English, the speakers are non-natives, yet highly experienced in English speaking. Voice data is recorded by two microphones, one dynamic passive headset for close talk (CT) and one far-field active condenser microphone for distant talk (DT) situated 50 cm in front of the subject. The mixed CT and DT signal at equal levels will be referred to as (MC) in the ongoing. Annotation was carried out by four male labelers after pre-segmentation into sub-speaker-turns by one annotator. The following table provides an overview of the per-LOI-contained samples considering only such with at least 3 out of 4 inter-labeler-agreement (ILA). The '?' indicates phrases that could not be assigned at this ILA level.

Table 1: Sub-turn LOI distribution in the database

LOI	0	1	2	?
[#]	319	1,762	171	1,677
[%]	8.1	44.8	4.4	42.7

As can be seen within the table, LOI 1 is clearly the pre-dominant LOI in this experiment. This may be due to the fact that more speech interaction occurs in the case of general interest. The next table shows the distribution of phrases based on their ILA in total and relative number.

Table 2: Inter-labeler agreement in the database

ILA	4/4	3/4	>2/4
[#]	580	1,672	**2,252**
[%]	14.8	42.6	**57.3**

These figures demonstrate that only around half of the 3,929 collected sub-speaker-turns are assignable by a minimum ILA of 3 out of 4. As a mean overall ILA on these samples 81.4% can be named. This may be seen as guideline for comparison throughout the up-coming accuracies by automatic classification. Additionally the spoken content and nonverbal interjections have been labeled. These interjections are *breathing*, *confirmation, coughing, hesitation, laughter*, *long pause, short pause* and other *human noise*. This additional labeling effort shall demonstrate the potential of such events within higher semantic analysis.

3. Acoustic Features

In respect of the quasi-stationary nature of a speech signal, firstly a pre-processing by windowing the signal with a Hamming-window function is fulfilled. The signal of interest is likewise split into successive 20 ms frames, windowed every 10 ms. In order to obtain a better representation in view of LOI content, feature contours containing information about intonation, intensity, harmonic structure, formants, and spectral development and shape are extracted. In detail these are: pitch based on time-domain calculation by auto-correlation function (ACF), window-function normalization and Dynamic Programming (DP) for global cost minimization, energy by frame-based signal-energy computation, formants' 1-5 amplitude, bandwidth, and frequency based on 18 LPC spectrum and DP, Mel-Frequency-Cepstral-Coefficients (MFCC) 1-16, spectral flux, 47 semi-tone-band interval emphasis and harmonic characteristic based on 1024-point DFT-spectrum, Harmonics-to-Noise Ratio (HNR) based on ACF in the time-domain, window-function normalization, shimmer and jitter of periodic parts, and 19 VOC19-coefficients [3]. Secondly, the derivation of speed and acceleration regression coefficients based on these Low-Level-Descriptors (LLD) is fulfilled as further information.

By LLD analysis a classification by means of dynamic modeling is already feasible. Yet, basing on our past experience [8] and in accordance with the common practice in the field [1] [5], we decided for a further processing step: In a third stage, higher-level functionals f are derived by means of descriptive statistics in order to project the multivariate time-series F on a static feature vector [4] and thereby become less dependent of the spoken phonetic content:

$$f : F \rightarrow \mathbb{R} \tag{1}$$

A systematic generation by calculation of moments, extreme values, and further shape characteristics out of each time series on a phrase basis leads to more than 5k features aiming at broad coverage of prosodic, articulatory and speech quality attributes. In detail the 18 functionals are: extreme values, extreme value positions, range, mean, centroid, standard deviation, quartiles, quartile ranges, 95% roll-off-point, Kurtosis, Skewness, and zero-crossing-rate. Thereby the number of features presented in [12] is consequently enlarged. The idea thereby is not to extract all these features for the actual LOI-detection, but to form a broad basis for self-learning feature-space optimization.

4. Linguistic Features

Beyond the analysis of acoustic properties of a speech signal, also the spoken content may carry cues in respect of a speaker's interest, and the combination of both analyses could be shown highly effective in our past related works in the field of speech emotion recognition. [9][10].

The precondition of linguistic analysis is to obtain the spoken content out of an audio-file. Within this work once manual annotations have been employed to obtain an impression of performance under idealistic speech recognition conditions and once a state-of-the-art MFCC and HMM-based tri-phone Large-Vocabulary Automatic Speech Recognition (ASR) engine was used.

For linguistic analysis a vector-space-representation popular in the field of document retrieval known as Bag-of-Words (BOW) has been chosen [2]. The motivation here fore is the effective fusibility of obtained linguistic features within the acoustic features on an early level [10]. Likewise loss of information is postponed to the final decision process allowing for the utmost decision basis. A term w_i within a phrase $\mathcal{S}=\{w_1,\dots,w_j,\dots,w_S\}$ is thereby projected onto a numeric attribute $x_i : w_i \rightarrow x_i$. The precondition is to establish a vocabulary $\mathcal{V}=\{w_1,\dots,w_i,\dots,w_V\}$ of terms of interest. In a first approach these are all different terms contained in the annotation of the data-set of interest. Throughout feature extraction a value for each term in $\mathcal{V}$ is calculated: Either 0 in case of no occurrence in the actual phrase, or 1 in case of a binary attribute's type, respectively its term frequency of occurrence (TF) for common BOW representation. A number of further refinement approaches exist as normalization to the phrase length, the inverse frequency of occurrence in the data-set known as Inverse Document Frequency (IDF), or logarithmic transform (log) to compensate linearity. Thereby an offset-constant $c=0.5$ is chosen, as many zero-occurrence cases will be observed. Our final per-term feature is calculated as follows and proved superior throughout evaluation to the named alternatives:

$$x_{\log TF,i} = \log\left(c + \frac{TF(w_i,\mathcal{S})}{|\mathcal{S}|}\right) \tag{2}$$

A drawback of this modeling technique is the lack of word order consideration. Still, great flexibility is obtained in comparison to e.g. class based N-Grams.

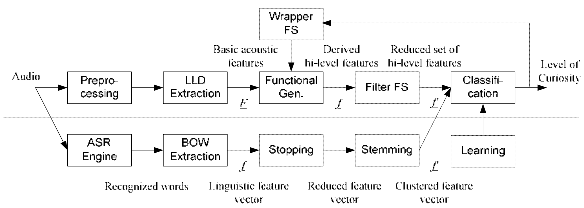
In general vocabularies will show too high a dimensionality (>1k terms) and contain many redundancy in view of the aimed at LOI-detection. Similar to acoustic feature reduction as described in the next section two standard techniques in linguistic analysis are therefore employed to reduce complexity: stopping and stemming. The first method directly reduces the vocabulary by eliminating terms of low relevance. This is realized based on Shannon's information as described in the ongoing. Stemming on the other hand clusters morphological variants of terms belonging to the same lexeme, i.e. having the same stem. Thereby the hit-rate of such clusters is directly boosted while reducing complexity at the same time. However, danger of over-stemming exists, i.e. clustering of terms that possess different meanings in view of LOI. We decided for an iterated Lovins-Stemmer, here fore, which bases on context-sensitive longest match stemming – a slight enhancement of the very traditional approach to stemming.

5. Optimization and Classification

In order to save extraction effort and reduce complexity throughout the succeeding classification process, features of high individual information are pre-selected by fast Information Gain Ratio (IGR) calculation and feature selection (FS) of attributes with high IGR (IGR-FS). This method bases on Shannon's information and is suited to find features of high individual relevance. Yet, redundancy of single high effectiveness is not filtered thereby. Within the reduced set by elimination of all zero information features, application of Sequential Forward Floating Search (SFFS) [6] leads to an optimal set as a whole and the overall minimum number of features [10]. SFFS employs a classifier, ideally the target one, as optimization criterion. Herein, powerful Support-Vector-Machines (SVM) is used to ensure high quality throughout selection (SVM-SFFS). SFFS is a Hill-Climbing search, and allows for forward and backward search

steps in order to cope with nesting effects. A search function is needed, as exhaustive search becomes NP-hard having such high dimensionality.
In view of optimal classification a variety of suited machine learning techniques is considered: instance based classifiers (1NN, kNN), Bayesian learners (NB, BN), Decision Trees (DT C4.5), and Support Vector Machines (SVM) [13]. Further more the construction of ensembles by AdaBoosting, Bagging, MultiBoosting, Voting and StackingC [11] to raise the performance of single classifiers and combine the strengths of diverse such in a synergetic way is tested to respect biases due to data sparseness typical in this field [10]. The following figure provides an overview of the recognition flow so far:

Figure 1: Overview Recognition of Interest



6. Experimental Results

As a general mean of evaluation, *10*-fold stratified cross validation (SCV) is employed. Thereby mean accuracies of correctly assigned samples throughout randomly shuffled but stratified and disjunctive cycles are provided. Alternatively speaker-independent accuracies are reported. In the ongoing results once for acoustic speech analysis, and once for linguistic such will be shown. Finally, the overall result by fusion of both information sources is shown.

Table 3: Classifier comparison, Top 82 SFFS acoustic features after IGR pre-selection, MC, 10-fold SCV

Accuracy [%]	**LOI 0,1,2**	**LOI 0 vs. 2**
1NN	76.7	88.8
kNN	81.6	90.8
DT C4.5	72.4	84.6
AdaBoosting C4.5	79.2	88.5
NB	51.9	79.9
BN	72.1	82.7
SVM	80.5	93.0
StackingC MLR Top 3	**81.7**	92.5
StackingC MLR Top 2	80.7	**92.7**

Table 3 shows results for the classifier and ensemble comparison within a 10-fold SCV on the named database with acoustic features only. Two different settings have been considered: discrimination of all three LOI levels, and only of LOI 0 vs. LOI 2. The latter scenario has been chosen, as the dataset is heavily unbalanced having all LOI included, but LOI 0 and LOI 2 are contained in a reasonable balance. At the same time this setting demonstrates the effectiveness of discrimination of neutral and highly interested phrases. The feature-set has been pre-selected and finally reduced in the named manner by successive IGR-FS and SVM-SFFS resulting in top 82 attributes for maximum performance and effectiveness. As can be seen in the table, a discrimination of all LOI is realizable in the same region as ILA on the set at 81.7% accuracy. However, discrimination of the two named LOI can be fulfilled with impressive 92.7% accuracy.
Considering speaker independency 79.5% mean accuracy are observed when splitting speakers into two halves and 2-fold cross-validating with optimal classifier configuration for full-blown LOI recognition applying feature reduction to 100 features by IGR-FS. While this is a speaker independent result, it is not the highest obtainable speaker independent accuracy, as more speakers could have been used for training and the features might have been further optimized by SVM SFFS. Yet it already demonstrates that SCV rates are close to speaker independent performance. Evaluating LOI 0 vs. LOI 2 exclusively a leave-one-speaker-out (LOSO) evaluation was carried out. Thereby 90.1% mean accuracy were observed when randomly selecting 10 speakers for evaluation and having SVM as classifier, IGR-FS set reduction to 300 features and subsequent SVM-SFFS reduction to top 135 features. The minimum accuracy observed thereby was 84.2%, the maximum 96.0% for a specific speaker, each.
Considering microphone set-up another test was carried out in a 10-fold SCV, comparing close-, distant-talk, and mixed channel. Features have been reduced by IGR-FS to 1k. As classifier SVM were chosen with a polynomial Kernel-function. 88.1% mean accuracy is observed for CT, 87.8% for DT, and 87.6% for MC within the discrimination of LOI 0 vs. LOI 2. Likewise rather insignificant influence of microphone positioning in the database can be named.
Within linguistic experiments test-runs employing an actual ASR and annotation based runs have been fulfilled. Firstly, a look at the minimum term frequencies within the set clearly speaks for problems arising when using a real ASR engine:

Table 4: Term numbers at diverse minimum TF levels, annotation-based (left) and ASR-based (right)

Min. TF	Annotation Terms [#]	ASR Terms [#]
1	1,485	1,568
2	645	351
3	422	191
4	336	136
5	277	109
10	149	51
20	98	20
50	48	8

The table shows more terms of single occurrence than actually contained in the vocabulary when using real ASR. This comes, as words are partly misrecognized and matched on diverse further terms. On the other hand side this diffusion by word errors also leads to fewer observations of the same terms: Already at a minimum TF of 2 within the database the annotation based level overtakes. Yet, BOW relies on high TF within a data-set. This can partly be repaired by stemming; assuming that phonetic mismatches lead to confusions within a lexeme.
Table 5 shows the 18 most relevant lexemes after iterated Lovins stemming and IGR-FS stopping. The final vocabulary size thereby is 639 lexemes instead of 1,485 terms. Using linguistic features only, maximum mean accuracy within 10-fold SCV, optimal feature type and SVM reached 79.4% for

the full blown LOI analysis based on annotation and 84.2% for discrimination of LOI 0 and LOI 2. Using ASR a drop to 69.8% is observed for LOI 0-2. 29.1% of the phrases led to no ASR output, as these sub-speaker-turns only consist of interjections or are too short. One of the main differences of annotation versus ASR thereby is the included annotation of non-verbal interjections or events as described within the database section. While the table above shows the high ranking of four of these events (in italics, as described in database section) on the ranks 1, 2, 8, and 9, a reduction of all such only led to an absolute accuracy drop of 1.9% having the same setup as described earlier: 10-fold SCV, SVM and LOI 0-2. Still, it might be of interest considering their automatic recognition within future work.

Table 5: Top 18 lexemes after stemming and IGR-FS based stopping. Stems are marked by *

Rank	Stem	IGR	Rank	Stem	IGR
1	*cough.*	0.2995	10	a	0.0308
2	*laugh.*	0.1942	11	that	0.0305
3	yeah	0.0514	12	car	0.0275
4	oh	0.0474	13	*hav	0.0263
5	*ver	0.0358	14	is	0.0258
6	if	0.0358	15	I	0.0252
7	*th	0.0337	16	*s	0.0230
8	*h.noise*	0.0325	17	and	0.0219
9	*hesit.*	0.0323	18	it	0.0219

The final table depicts the gain achieved by fusion of acoustic and linguistic features on the spontaneous database. SVM proved the best choice here. However, only a slight improvement is obtained compared to the extra need of an ASR unit.

Table 6: Fusion of linguistic features and Top 82 SFFS acoustic features after IGR pre-selection, MC, 10-fold SCV

Accuracy [%]	**LOI 0,1,2**	**LOI 0 vs. 2**
SVM	**82.8**	**93.6**

The F_1-measure for LOI 0 thereby was 95.4% and 89.3% for LOI 2.

7. Conclusions

At this point a number of conclusions shall be derived: Firstly, acoustic three-level interest recognition reached the inter-labeler-agreement with 81% accuracy within 10-fold cross-validation and 80% within 2-fold speaker independent evaluation. Discrimination of neutrality and high interest can be fulfilled with impressive 93% accuracy within 10-fold cross-validation and 90% within speaker-independent evaluation. In this test low microphone positioning dependency can be reported. Considering linguistic analysis it can be stated that it leads to reasonable results as well: 79% for full-blown interest detection, 84% for neutrality vs. high interest, but a performance drop with genuine ASR occurs probably due to non-native speakers, affective speaking style influences, many very short phrases (~30%) and lack of interjection recognition. Still, the best result is obtained by fusion of acoustic and linguistic features reaching 82.8% respectively 93.6% for the discrimination of two or three levels of interest. Overall, outstanding performances for the recognition of spontaneous interest in human communication can be reported. Future works will aim at ASR refinement, word-class frequencies and further linguistic features, inclusion of word- or syllable levels, and context modeling of general LOI development.

8. References

[1] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G.: Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Technical LS-8 Report 23*, Dortmund, 1997.

[3] Holmes, J. N.: The JSRU 19-channel Vocoder. *IEE Proceedings*, vol. 1, part F, pp. 127, 1980.

[4] Mierswa, I.: Automatic Feature Extraction from Large Time Series. *Proceedings of the 28. Annual Conference of the GfKl 2004*, Springer, pp. 600-607, 2004.

[5] Pantic, M; Rothkrantz, L..: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE*, Vol. 91, pp. 1370-1390, Sep. 2003.

[6] Pudil, P.; Novovičová, J.; Kittler, J.: Floating search methods in feature selection, *Pattern Recognition Letters*, Vol. 15/11, pp. 1119–1125, Nov. 1994.

[7] Qvarfordt, P.; Beymer, D. Zhai, S.: RealTourist – A Study of Augmenting Human-Human and Human-Computer Dialogue with Eye-Gaze Overlay, *INTERACT 2005*, LNCS 3585, pp. 767 – 780, 2005.

[8] Schuller, B.; Rigoll, G.; Lang, M.: Hidden Markov Model-Based Speech Emotion Recognition. *Proceedings of the ICASSP 2003*, Vol. II, pp. 1-4, Hong Kong, China, 2003.

[9] Schuller, B.; Ablaßmeier, M.; Müller, R.; Reifinger, S.; Poitschke, T.; Rigoll, G.: *"Speech Communication and Multimodal Interfaces",* in Advanced Man Machine Interaction, K.-F. Kraiss (ed.), Springer, Berlin, Heidelberg, ISBN: 3-540-30618-8, pp. 141-190, 2006.

[10] Schuller, B.; Müller, R.; Lang, M.; Rigoll, G.: Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. *Proceedings of the INTERSPEECH 2005*, Lisbon, Portugal, pp. 805-809, 04.-08.09.2005.

[11] Seewald, A. *Towards understanding stacking – Studies of a general ensemble learning scheme*. PhD-Thesis, TU Wien, 2003.

[12] Vogt, T.; Andre, E.: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. *Proceedings of the ICME 2005*, Amsterdam, Netherlands, 2005.

[13] Witten, I. H.; Frank, E.: *Data Mining, Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, pp. 133, 2000.