Temporal and Situational Context Modeling for Improved Dominance Recognition in Meetings

Martin Wöllmer, Florian Eyben, Björn Schuller, Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Germany
[woellmer,eyben,schuller,rigoll]@tum.de

Abstract

We present and evaluate a novel approach towards automatically detecting a speaker's level of dominance in a meeting scenario. Since previous studies reveal that audio appears to be the most important modality for dominance recognition, we focus on the analysis of the speech signals recorded in multiparty meetings. Unlike recently published techniques which concentrate on frame-level hidden Markov modeling, we propose a recognition framework operating on segmental data and investigate context modeling on three different levels to explore possible performance gains. First, we apply a set of statistical functionals to capture large-scale feature-level context within a speech segment. Second, we consider bidirectional Long Short-Term Memory recurrent neural networks for long-range temporal context modeling between segments. Finally, we evaluate the benefit of situational context incorporation by simultaneously modeling speech of all meeting participants. Overall, our approach leads to a remarkable increase of recognition accuracy when compared to hidden Markov modeling.

Index Terms: dominance recognition, meeting analysis, Long Short-Term Memory, audio feature extraction

1. Introduction

Face-to-face meeting scenarios have attracted a lot of attention within the research fields of speech recognition, affective computing, and social signal processing [1]. The common goal of pattern recognition systems tailored for meeting data is to make meetings more efficient, e. g., by automatically summarizing them via generation of transcriptions or extraction of important events (decision making), or by applying machine learning for creating virtual conference directors and meeting browsers [2]. An essential indicator for characterizing the course of a meeting is the level of dominance of the individual participants. Usually a certain order of dominance is established after a short period of time, even if the participants do not know each other [3]. Typically, the dominance expressed by the individuals varies over time and carries important information about related social signals and attributes such as activity or hierarchical ranking and about the trait of the meeting as a whole.

Automatically recognizing a speaker's level of dominance has been attempted in various studies, exploiting different modalities. One possibility is to use high-level features such as speech transcriptions [4] which, however, implies high latency and a high real-time factor. Other systems focus on evaluating the speaking length of each speaker in a segment via speaker diarization [5] and on applying high-level audio-visual cues [6]. In [3], the authors evaluate both, low-level audio and video features as well as high level semantic features for dominance recognition in meeting rooms. Experiments on the AMI corpus [7] show that the best accuracy can be obtained with (low-level) audio features only.

Based on these findings, this paper focuses on speech-based dominance detection and demonstrates how accuracy can be enhanced by appropriate context modeling. Unlike the system introduced in [3] which – as it is common practice in speech recognition – models framewise Mel-Frequency Cepstral Coefficients (MFCC) and their derivatives via Hidden Markov Models (HMM), our technique builds on recent advances in emotion recognition and employs a large set of spectral, prosodic, and voice quality low-level descriptors (LLD). All features are extracted in real-time using our open-source toolkit for large-scale speech feature extraction openSMILE [8]. To capture feature-level context and dynamics we apply a set of statistical functionals on the LLDs which results in one high-dimensional feature vector per speech segment. Contextual information between successive speech segments is accounted for by employing Long Short-Term Memory (LSTM) networks which are known to be well suited for affective computing [9] as their model architecture allows for temporal long-range context exploitation. Motivated by studies as in [10] where it was shown that emotion recognition profits from taking into account speech cues from a speaker's interlocutor, we also consider situational context in the sense of other participants' speech by simultaneously modeling multiple speakers. As in [3], we evaluate our techniques on a subset of the AMI corpus and achieve a considerable accuracy gain compared to previous methods based on HMMs.

2. Database

Experimental results are based on the same subset of the AMI corpus [7] as used in [3]. It consists of 36 meeting recordings, each having a length of five minutes, which results in a total length of 180 minutes. In each meeting four participants are located in the IDIAP smart meeting room, equipped with 22 microphones and seven cameras. As in [3], we exclusively consider the four close talking microphones for speech-based dominance recognition. Thus, the audio material used for this study has a total length of 12 hours. Since the speech signals captured by the close talking microphones contain some cross-talk in case multiple speakers speak at the same time, an additional cross-talk free version of the corpus was created by manually removing of cross-talk. The aim was to investigate both, a realistic scenario including distortions and a scenario with perfect cross-talk cancellation.

For dominance annotation and recognition, segments of ten seconds each have been created, i. e., each ten second fragment of each close talking microphone recording had been given a dominance label by two annotators. The original annotated levels of dominance range from 1 to 5, meaning from 'absent' to

Table 1: *31 low-level descriptors (LLD).*

Energy & Spectral (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, entropy, spectral variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, MFCC 1-10
Voicing related (6)
F_0 (Sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), logarithmic Harmonics-to-Noise Ratio (logHNR)

'extremely dominant', however, as only 12 of the 4 320 segments in the database are labeled with dominance level 5, we clustered together levels 4 and 5. For details on the annotation and the inter-labeler agreement the reader is referred to [3].

In conformance with [3] we apply a nine-fold cross validation with speaker disjoint training, validation, and test sets. For each fold we divide the database into 28 meetings for training, four meetings for validation, and four meetings for testing. This corresponds to 3 360, 480, and 480 segments, respectively. Note that the validation set is exclusively used for determining the optimum number of epochs for neural network training.

3. Feature Extraction

In contrast to the system proposed in [3], which directly processes low-level MFCC features via HMMs, our approach is based on a large set of LLDs and derivatives of LLD combined with suited statistical functionals to capture speech dynamics within a segment. All features and functionals are computed using our online audio analysis toolkit openSMILE [8]. The audio feature set consists of 1 941 features, composed of 25 energy and spectral related low-level descriptors x 42 functionals, 6 voicing related LLD x 32 functionals, 25 delta coefficients of the energy/spectral LLD x 23 functionals, 6 delta coefficients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced durational features. Details on the LLD and functionals are given in Tables 1 and 2, respectively. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition. The functional set has been based on similar sets, such as the one used for the Interspeech 2011 Speaker State Challenge [11], but has been carefully reduced to avoid LLD/functional combinations that produce values which are constant, contain very little information and/or a high amount of noise. Since the speakers' levels of dominance are annotated every ten seconds (see Section 2), our speech feature extractor uses 10 s fragments of speech for the calculation of statistical functionals which then are processed by the dominance classification back-end.

In order to reduce the size of the resulting feature space, we considered a cyclic Correlation based Feature Subset Selection (CFS) using the training set of each fold. This results in an automatic selection of between 47 and 86 features, depending on the cross-talk scenario and on the fold.

Table 2: *Set of all 42 functionals. [1]not applied to delta coefficient contours. [2]for delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]not applied to voicing related LLD.*

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean standard deviation, flatness skewness, kurtosis quartiles, and inter-quartile ranges 1 %, 99 % percentile percentile range 1 %–99 % percentage of frames contour is above: min + 25%, 50%, and 90 % of the range percentage of frames contour is rising max, mean, min segment length[3] standard deviation of segment length[3]
Regression functionals[1] (4)
linear regression slope, and approximation error (linear), quadratic regression coefficient a, and approx. error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances amplitude mean of maxima amplitude mean of minima amplitude range of maxima
Other[1,3] (6)
LP gain, LPC 1-5

4. Classification

Most systems which directly process framewise low-level features such as MFCCs employ some form of dynamic Bayesian network for decoding. The predominant methodology used in LLD-based speech and emotion recognition is to apply Hidden Markov Models capturing short-term context via state transition likelihoods. However, recent studies show that for affective computing turnwise or segmental modeling tends to prevail over processing frame by frame observations with HMMs [12]. Widely used classifiers operating on static feature vectors are, e. g., Support Vector Machines (SVM) or Multilayer Perceptrons (MLP). To exploit context between successive speech segments for improved dominance recognition, this study considers *recurrent* neural network (RNN) architectures which take into account past observations by cyclic connections in the network's hidden layer. For off-line sequence labeling problems, also *future* context can be modeled via *bidirectional* RNNs (BRNN). Bidirectional networks have access to both, past and future observations by applying two hidden layers, one for forward processing and one for backward processing. These two hidden layers are connected to the same output layer (see [13] for details). For dominance recognition BRNNs can be employed whenever the real-time constraint can be relaxed, i. e., when focusing on off-line processing or when a short latency is tolerable, so that the system can be operated with a look-ahead buffer.

In our experiments we also investigate a more advanced technique for neural network based context modeling. It is based on the Long Short-Term Memory principle originally introduced in [14]. LSTM networks use so-called memory blocks instead of conventional hidden cells which allows them to access and model a self-learned amount of long-range temporal context. Each memory block consists of one or more memory cells and multiplicative input, output, and forget gates. The cell input is scaled by the activation of the input gate, the output by the activation of the output gate, and the previous cell value by the activation of the forget gate. Thus, the network can perform read, write, and reset operations, and – unlike traditional RNNs which are affected by the *vanishing gradient problem* – has access to an arbitrary amount of context information. LSTM networks have shown remarkable performance in a variety of pattern recognition tasks, including handwriting recognition [13], speech recognition, and affective computing [9, 12]. Details on the LSTM technique and on its bidirectional extension (BLSTM) can be found in [13].

5. Experiments

5.1. Experimental Settings

All dominance recognition experiments were conducted using a cyclic nine-fold cross validation on the AMI meeting database (see Section 2). We investigated recognition performance on data with and without cross-talk cancellation considering either the full set of 1 941 features per segment or the reduced set obtained via feature selection as detailed in Section 3. Five different classification approaches were tested; each of them operating on segmental data: Support Vector Machines, RNNs, bidirectional RNNs, LSTM networks, and bidirectional LSTMs. The RNNs and LSTM networks consist of 128 hidden cells and memory blocks, respectively. Each memory block contains one memory cell. The number of input nodes corresponds to the number of different features per speech segment whereas the number of output nodes corresponds to the number of target classes, i. e., we used four output nodes representing the four levels of dominance. All networks were trained using a learning rate of 10^{-5}. The bidirectional networks consist of two hidden layers (one for forward and one for backward processing) with 128 cells / memory blocks per input direction. As abort criterion for training we evaluated the classification performance on the validation set of the respective fold. The applied SVMs have a polynomial kernel (degree 1) and are trained using the sequential minimal optimization (SMO) algorithm.

5.2. Multiple Speaker Modeling

Aiming to improve the recognition of a single speaker's level of dominance by situational context modeling, i. e., by modeling multiple speakers at the same time, we investigated the effect of extending the feature vectors so that they contain speech features from the speaker whose level of dominance we want to classify, as well as the speech features from the remaining three meeting participants. The motivation for this is to exploit potential negative correlations between the dominance level of the considered speaker and his or her interlocutors, assuming that not all participants will have a high level of dominance at the same time. A classifier processing the speech features of all four participants at the same time will thus have to learn correlations between the first fourth of the extended feature vector and the ground truth dominance label (corresponding to speech from the speaker under consideration) as well as corresponding inverse correlations between the remaining three-fourths of the feature vector and the dominance label.

5.3. Results and Discussion

Table 3 shows the dominance recognition performance for the different classification approaches and the effect of cross-talk cancellation, feature selection and multiple speaker modeling. Since classes are heavily unbalanced, we decided for the F1-measure (harmonic mean between unweighted recall and unweighted precision) as evaluation criterion. When considering the best result reported in [3] for an HMM-based system processing low-level MFCC features (recognition rate of 54.90 %), we see that the most remarkable performance gain results from modeling segmental statistical functionals of a large set of LLDs (see Section 3) instead of framewise MFCCs: The best systems proposed in this study achieve an F1-measure of 62.93 % for data including cross talk and 65.25 % for cross-talk free data, corresponding to a recognition rate (or *weighted* accuracy) of 68.17 % and 70.26 %, respectively. Thus, modeling feature-level context within a speech segment via statistical functionals of large-scale LLDs as used in modern emotion recognition systems [9, 11] leads to a gain in dominance recognition rate of up to 13.3 % absolute, considering data without cross-talk cancellation.

Depending on the size of the feature space, context modeling between successive speech segments can also lead to enhanced recognition performance. When considering single speaker modeling without feature selection, we see that there is no clear trend as far as the F1-measures of the different classifiers is concerned – apart from SVMs operating on cross-talk free data (62.53 %), which perform worse than RNN-based approaches exploiting context. However, when using CFS feature selection, which generally increases performance for most scenarios, we observe that context exploitation can increase performance: For the realistic cross-talk setting SVMs achieve a comparably low F1-measure of 55.64 % which can be enhanced via long-range context modeling by bidirectional LSTMs (62.93 %). Similar trends can be observed for the cross-talk free scenario, where F1-measure increases from 63.67 % to 65.25 % when replacing SVMs with unidirectional LSTM networks. For most experimental settings, there is no significant difference between results obtained with unidirectional and bidirectional processing which means that the real-time version of the proposed dominance recognition system does not imply lower performance. Applying multiple speaker modeling in combination with the full feature set does not improve performance, as the resulting feature vector dimension of four times 1 941 seems to be too large for the given amount of training data. Yet, when applying feature selection and focusing on the cross-talk scenario, situational context increases the performance of most classifiers, with the most remarkable enhancement from 55.64 % to 61.30 % for SVM-based recognition. Multi-speaker context especially helps when having to process data that includes cross-talk – probably because classifiers seem to learn to distinguish cross-talk from speech when features from all speakers are clustered together in one feature vector. Performance gains by temporal and situational context modeling seem to be not fully complementary since classifiers ignoring temporal segment-level context profit more from multiple speaker modeling than, e. g., RNN-based or LSTM-based recognizers.

To further study the potential of multiple speaker modeling for dominance recognition we trained multi-task BLSTM networks for predicting the level of dominance of all four speakers at the same time, using features from all speakers, as in our previous multi speaker modeling experiments. Combined

Table 3: *Dominance recognition results in terms of F1-measure on the test set, averaged over nine folds: results for different classifiers, with and without cross-talk cancellation, with and without feature selection via CFS, and with and without multiple speaker modeling.*

classifier	cross-talk cancellation	*without feature selection*		*with feature selection*	
		single speaker modeling	multiple speaker modeling	single speaker modeling	multiple speaker modeling
BLSTM	✓	64.29	63.05	64.69	62.42
LSTM	✓	64.82	64.45	**65.25**	63.89
BRNN	✓	63.63	63.37	63.24	62.88
RNN	✓	65.24	63.49	64.82	64.52
SVM	✓	62.53	59.00	63.67	64.40
BLSTM	✗	60.44	59.01	**62.93**	61.31
LSTM	✗	61.42	58.95	61.64	62.65
BRNN	✗	60.62	60.77	60.53	61.52
RNN	✗	61.57	61.17	60.51	61.92
SVM	✗	61.27	58.72	55.64	61.30

with feature selection, this led to a decreased F1-measure of 59.08 % compared to 62.93 % for single speaker modeling and data with cross-talk. Also the inclusion of an oracle feature encoding the ground truth level of dominance of the three other speakers did not improve the F1-measure for this experimental setting (62.54 %). Correlating the ground truth dominance labels of one speaker with the labels of the speaker's interlocutors leads to a very small negative cross-correlation of -0.16 on average, which is a further indicator for the rather limited gain of multiple speaker modeling, compared to other forms of context usage such as temporal context exploitation and modeling of feature-level dynamics.

6. Conclusion

We introduced a novel technique for recognizing a speaker's level of dominance in a meeting scenario. The system processes segmental speech features obtained by our toolkit for large-scale feature extraction (openSMILE [8]). Compared to previously introduced HMM-based systems using standard MFCC features [3] our approach enables more accurate dominance estimation which can be attributed to the combination of segmental processing and context modeling within and between speech segments. We found that temporal context exploitation via RNNs or Long Short-Term Memory networks tends to prevail over SVM-based recognition. As the levels of dominance of the individual speakers at a given time are almost fully uncorrelated, the gain of modeling situational context in the form of observing multiple speaker's features at the same time, is comparably small and rather partly compensates performance degradations by cross-talk.

In the future, we intend to combine the proposed segmental feature extraction scheme with useful speech-based high-level features (e. g., as applied in [4]). Further, it seems promising to integrate our dominance detector into systems which are tailored for processing room microphone recordings and thus include modules for source separation and dereverberation.

7. Acknowledgements

The authors would like to thank Benedikt Hörnler, Emanuele Principi, and Stefano Squartini for providing the database used in this paper.

8. References

[1] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, D. van Leeuwen, M. Lincoln, and V. Wan, "The 2007 ami(da) system for meeting transcription," in *Multimodal Technologies for Perception of Humans*. Springer, 2009, pp. 414–428.

[2] M. Al-Hames, B. Hörnler, R. Müller, J. Schenk, and G. Rigoll, "Automatic multi-modal meeting camera selection for video-conferences and meeting browsing," in *Proc. of ICME*, Beijing, China, 2007, pp. 2074–2077.

[3] B. Hörnler and G. Rigoll, "Multi-modal activity and dominance detection in smart meeting rooms," in *Proc. of ICASSP*, Taipei, Taiwan, 2009, pp. 1777–1780.

[4] R. J. Rienks and D. Heylen, "Automatic dominance detection in meetings using easily obtainable features," in *Proc. of MLMI*, Edinburgh, Scotland, 2006, pp. 76–86.

[5] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 847–860, 2011.

[6] O. Aran and D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in group conversations," in *Proc. of ICPR*, Istanbul, Turkey, 2010, pp. 3687–3690.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: a pre-announcement," in *Proc. of MLMI*, 2006, pp. 28–39.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[9] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[10] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. of Interspeech*, Brighton, UK, 2009, pp. 1983–1986.

[11] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in *Proc. of Interspeech 2011*, Florence, Italy, 2011, pp. 3201–3204.

[12] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 2362–2365.

[13] A. Graves, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Technische Universität München, 2008.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.