Dominance detection in a reverberated acoustic scenario

Emanuele Principi, Rudy Rotili, Martin Wöllmer, Stefano Squartini, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Principi, Emanuele, Rudy Rotili, Martin Wöllmer, Stefano Squartini, and Björn Schuller. 2012. "Dominance detection in a reverberated acoustic scenario." *Lecture Notes in Computer Science* 7367: 394–402. https://doi.org/10.1007/978-3-642-31346-2_45.

Dominance Detection in a Reverberated Acoustic Scenario

Emanuele Principi[1], Rudy Rotili[1], Martin Wöllmer[2], Stefano Squartini[1], and Björn Schuller[2]

[1] Dipartimento di Ingegneria dell'Informazione,
Università Politecnica delle Marche, Ancona, Italy
{e.principi,r.rotili,s.squartini}@univpm.it
[2] Institute for Human-Machine Communication,
Technische Universität München, Germany
{woellmer,schuller}@tum.de

Abstract. This work proposes a dominance detection framework operating in reverberated environments. The framework is composed of a speech enhancement front-end, which automatically reduces the distortions introduced by room reverberation in the speech signals, and a dominance detector, which processes the enhanced signals and estimates the most and least dominant person in a segment. The front-end is composed by three cooperating blocks: speaker diarization, room impulse responses identification and speech dereverberation. The dominance estimation algorithm is based on bidirectional Long Short-Term Memory networks which allow for context-sensitive activity classification from audio feature functionals extracted via the real-time speech feature extraction toolkit openSMILE. Experiments have been performed suitably reverberating the DOME dataset: the absolute accuracy improvement averaged over the addressed reverberated conditions is 32.68% in the most dominant person estimation task and 36.56% in the least dominant person estimation one, both with full agreement among annotators.

1 Introduction

Recently, a certain attention has been paid by the scientific community to the development of automatic systems for dominance detection in small-groups [1]. Information coming from speech, but also from gesture, posture and face movements, can be extracted from the meeting activity and then be processed by expert algorithms in order to automatically detect the participants' level of dominance.

Dominance can be defined in multiple ways: it is often related to the notion of power, i.e. "the capacity to produce intended effects, and in particular, the ability to influence the behaviour of another person" [1]. This leads to defining dominance as a set of "expressive, relationally based communicative acts by

which power is exerted and influence achieved", "on behavioural manifestation of the relational construct of power", and "necessarily manifest" [1]. Dominance is directly related to the participant activity level [1]: persons with higher vocal and visual activity (e.g. body movement and gestures correlated with speaking activity) are often perceived as more dominant [2].

Several approaches have been proposed in the literature to address the dominance detection task. In [3], dominance is estimated calculating the speaking length of each speaker in a segment by means of the ICSI speaker diarization system. The system is able to work in real-time, but not online since the speaker diarization stage operates on the entire signals. The authors performed experiments on a subset of the AMI corpus using single distant microphone and headset signals. Reverberation and additive noise have not been taken into account. The methods proposed in [2,4] combine high level audio and visual features. They detect dominance levels either with a rule-based estimator, or with a Support Vector Machine classifier. Experiments are performed on the DOME dataset as considered herein [5] using individual headset microphones. In [6], two solutions for audio-visual activity and dominance detection are proposed: in the first, detection is performed using low level features and classification through Hidden Markov Models (HMM). In the second, a higher level feature containing the information about the current status of the group is added, and a two layer HMM system is employed for classification. Similarly to [3], experiments are conducted on a subset of the AMI corpus, this time employing individual headset microphones only, and annotated with participants' activity levels.

This paper addresses dominance detection in reverberated environments. Multiple distant microphones are used to acquire voices of meeting participants and the presence of the reverberation effect is dealt with by means of a recently proposed speech enhancement front-end [7]. Here, other sources of degradation, such as additive noise, are not considered. The enhanced signals are processed by the dominance detector stage which estimates the most and least dominant person in a segment using nonverbal vocalic cues. The full system block-scheme is shown in Fig. 1. The performance of the proposed framework are evaluated suitably reverberating the DOME dataset [5] with three different reverberation times: the obtained results show that both in estimating the most and least dominant person, the proposed framework achieves accuracies close to the non-reverberated condition ones.

The paper outline is the following. Sec. 2 briefly describes the speech enhancement front-end. Sec. 3 details the algorithm developed for dominance estimation. Sec. 4 discusses the experimental setup and the performed experiments. Finally, in Sec. 5 conclusions are drawn and future developments are proposed.

2 Speech Enhancement Front-End

The objective of the speech enhancement front-end is recovering the original clean speech sources. This is performed by means of a "context-aware" speech dereverberation approach, which includes the automatic identification of who

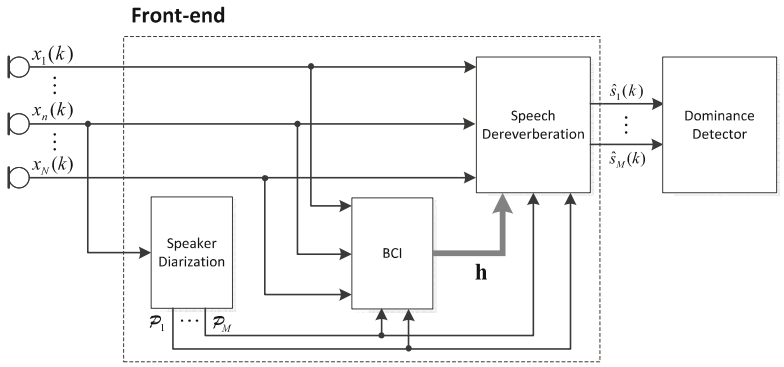


Fig. 1. Block diagram of the dominance detection framework

is speaking, the estimation of the unknown room IRs and the application of a knowledgeable dereverberation process to restore the original speech quality. To achieve such a goal, the framework proposed in [7] by some of the authors has been used. The framework consists of three stages: speaker diarization, blind channel identification and speech dereverberation.

Assuming M independent speech sources and N microphones, the relationship between them is described by an $M \times N$ MIMO FIR (Finite Impulse Response) system. According to such a model and denoting with $(\cdot)^T$ the transpose operator, the following equations (in the time and z domain) for the n-th microphone signal hold:

$$x_n(k) = \sum_{m=1}^{M} \mathbf{h}_{nm}^T \mathbf{s}_m(k, L_h), \qquad X_n(z) = \sum_{m=1}^{M} H_{nm}(z) S_m(z), \tag{1}$$

where $\mathbf{h}_{nm} = [h_{nm,0}\ h_{nm,1}\ \ldots\ h_{nm,L_h-1}]^T$ is the L_h-taps IR between the n-th microphone and m-th source $\mathbf{s}_m(k, L_h) = [s_m(k)\ s_m(k-1)\ \ldots\ s_m(k-L_h+1)]^T$, with $(m = 1, 2, ..., M,\ n = 1, 2, ..., N)$.

The speaker diarization stage drives the BCI and dereverberation blocks so that they can operate into speaker-homogeneous regions. The algorithm consists of two phases, training and recognition. In the first, 19 Mel-Frequency Cepstral Coefficients (MFCC) plus their first and second derivatives are obtained from the input signals. Cepstral mean normalization is applied to deal with stationary channel effects. Speaker models are represented by mixture of Gaussians trained by means of the expectation maximization algorithm. The end accuracy at convergence and the number of Gaussians have been empirically determined on meetings IS1004a-d of the AMI corpus and set respectively to 10^{-4} and 100. In the recognition phase, the input signal is divided into non overlapping chunks, and feature vectors are extracted as in the training phase. Participants' identities are then determined using majority vote on the likelihoods.

The blind channel identification stage is based on the so-called Unconstrained Normalized Multi-Channel Frequency domain Least Mean Square algorithm (UNMCFLMS) [8], a technique that represents an appropriate choice in terms of estimation quality and computational cost. Though UNMCFLMS allows the

estimation of long IRs, it requires a high input signal-to-noise ratio. Here, the noise free case has been assumed and future developments will consider improvements to make the algorithm more robust to the presence of noise.

The dereverberation stage is based on the Multi-channel Inverse Theorem (MINT) method. Given the SIMO system corresponding to source s_m, let us consider the polynomials $G_{s_m,n}(z), n = 1, 2, \ldots, N$ as the dereverberation filters to be applied to the SIMO outputs to provide the final estimation of the clean speech source s_m, according to the following:

$$\hat{S}_m(z) = \sum_{n=1}^{N} G_{s_m,n}(z) X_n(z). \tag{2}$$

The dereverberation filters can be obtained using the well known Bezout's Theorem. However, such a technique requires a matrix inversion that requires a high computational cost, especially in the case of long IRs. Therefore, in [7] the efficiency of the algorithm has been improved employing an adaptive approach.

3 Dominance Detector

Dominance detection is performed in three steps: in the first, feature vectors are extracted from the input signals every ten seconds. In the second, meeting participants' activity level is estimated by means of a Long Short-Term Memory network. In the third, the most and least dominant persons are estimated through a majority vote on the activities.

3.1 Speech Feature Extraction

For speech feature extraction, the online audio analysis toolkit openSMILE [9] is employed. We use the same set of 1 941 audio features as applied in [10]. It is composed of 25 energy and spectral related low-level descriptors (LLD) 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. 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, 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.

3.2 Most and Least Dominant Person Estimation Based on LSTM

Building on recent studies in the field of context-sensitive affective computing and human behaviour analysis [11], an activity classification framework that is based on bidirectional Long Short-Term Memory has been designed. The basic concept of Long Short-Term Memory (LSTM) networks was introduced in [12]

Table 1. Agreement statistics. "Full" indicates that three annotators agree, "Majority" indicates that two annotators agree, "None" indicates no agreement.

	Full	Majority	None
Most Dominant Person	58.62%	37.93%	3.45%
Least Dominant Person	53.45%	39.66%	6.89%

and can be seen as an extension of conventional recurrent neural networks that enables the modeling of long-range temporal context for improved sequence labeling. LSTM networks are able to store information in linear memory cells over a longer period of time and can learn the optimal amount of contextual information relevant for the classification task. An LSTM hidden layer is composed of multiple recurrently connected subnets (so-called *memory blocks*). Every memory block consists of self-connected *memory cells* and three multiplicative *gate* units (input, output, and forget gates). Since these gates allow for write, read, and reset operations within a memory block, an LSTM block can be interpreted as (differentiable) memory chip in a digital computer. Further details on the LSTM principle can be found in [13]. The most and least dominant persons in a meeting are estimated through a majority vote approach: the most (respectively, least) dominant person is the one that is classified as the most (respectively, least) active for the majority of segments.

4 Experiments

4.1 Corpus Description

Experiments have been conducted on the DOminance in MEetings dataset (DOME) [5], a subset of the AMI corpus [14] annotated with dominance levels. "Meeting Set 1" has been chosen in order to compare the obtained results with previous works on dominance estimation [2,3]. This set consists of 58 five minutes long segments extracted from 11 AMI scenario meetings. The total number of speakers is 20 and the female/male ratio is 42.86%. For each segment, dominance annotations have been performed by three annotators according to their level of perceived dominance. The distribution of agreement types is shown in Table 1.

Two main dominance tasks are defined in DOME: estimating the most dominant person and estimating the least dominant person. Based on the annotators' level of agreement, DOME defines four tasks:

- **FMD: F**ull agreement set, **M**ost **D**ominant person estimation task (34 segments).
- **FLD: F**ull agreement set, **L**east **D**ominant person estimation task (31 segments).
- **MMD: M**ajority agreement set, **M**ost **D**ominant person estimation task (56 segments).
- **MLD: M**ajority agreement set, **L**east **D**ominant person estimation task (54 segments).

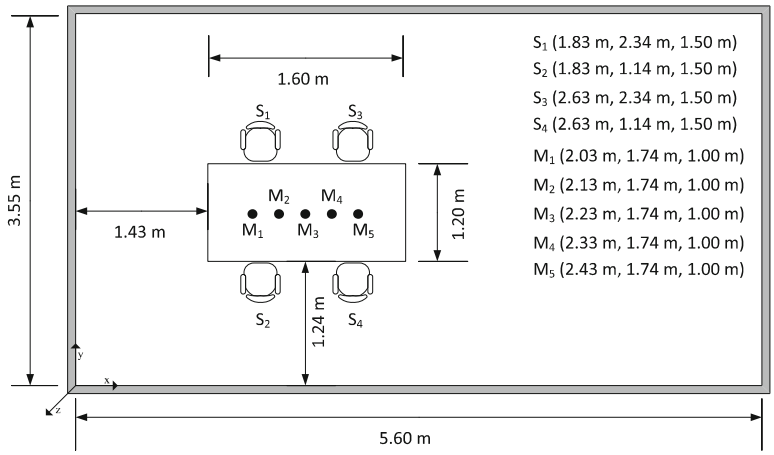


Fig. 2. Room setup: x, y and z coordinates are shown in brackets

4.2 Acoustic Scenario

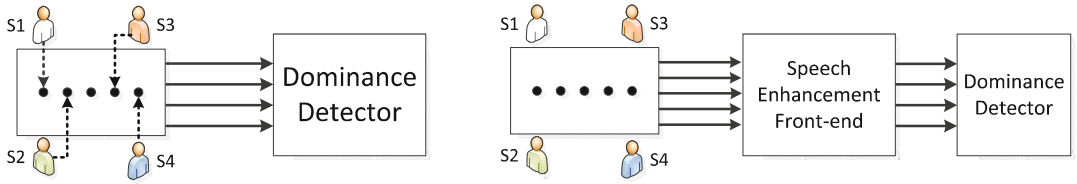
The scenario under study is shown in Fig. 2: an array of five microphones is placed at the centre of the meeting table and four speakers are sitting around it. The number of microphones has been chosen taking into account that it must be greater than the number of speakers [8]. The inter-microphone distance is 10 cm and represents a good comprise between impulse response diversification, which increases with the inter-microphone distance, and the need for a reasonably sized array. It is worth highlighting that the UNMCFLMS and MINT algorithms do not suffer from the spatial aliasing problem as delay and sum beamformer [15]. Microphone signals have been created by manually removing cross-talk from the headset sources and convolving them with impulse responses 1024 taps long. RIRs have been generated using Habets' RIR Generator tool[1], and represent three different reverberation times (T_{60}): 120 ms, 240 ms and 360 ms. Cross-talk free individual headset sources will be denoted as "Clean" in the following sections.

4.3 Dominance Detector Training and Evaluation Procedure

The networks used for the experiments consist of 1 941 input nodes (one for each speech feature extracted from 10 s of speech), 128 memory blocks containing one memory cell each, and four output nodes that represent the likelihoods of the four activity classes.

We trained a BLSTM network on the transcribed meeting segments used in [6], excluding segments that also occur in the DOME corpus. This results in a training database consisting of 26 meeting segments of five minutes each. As test set, we used the whole DOME corpus. All features were mean and variance normalized prior to processing via BLSTM networks. Means and variances were calculated from the training set only. During training a learning rate of 10^{-5}

[1] `http://home.tiscali.nl/ehabets/rirgenerator.html`



(a) *No front-end* configuration (b) *Full-system* configuration

Fig. 3. System configurations. In (a), the central microphone is not used and the dashed arrows denote a logical link between speakers and microphones.

and a momentum of 0.9 are used. Zero mean Gaussian noise with standard deviation 0.6 was added to the inputs in the training phase in order to improve generalization. Prior to training, all weights were randomly initialised in the range from -0.1 to 0.1. Input and output gates used tanh activation functions, while the forget gates had logistic activation functions. Due to the observed fast convergence, training was aborted after 10 epochs.

4.4 Results

The system evaluation has been conducted considering two configurations, "No front-end" and "Full-system". In the first, the speech enhancement front-end is not present and the dominance detector operates on four microphone signals (Fig. 3a). Each microphone is logically associated to a single speaker, meaning that the dominance detector expects each signal to contain only one voice. The purpose of this experiment is to highlight the need for a front-end able to divide and dereverberate the inputs. The "Full-system" configuration represents the proposed framework as shown in Fig. 3b and described in Sec. 2.

The dominance detection accuracies obtained on "Clean" signals are 85.29% (FMD), 80.65% (FLD), 76.79% (MMD) and 62.96% (MLD). Similar results have been obtained in [3], where the DOME dataset has been used as well. In the most dominant person estimation tasks accuracies are very similar: 85% in FMD and 77% in MMD. In the least dominant person estimation tasks, they report a higher value in FLD (84%) and a lower value in MLD (59%). It is worth pointing out that differently from [3], the system described here operates entirely online.

Table 2 shows the dominance detection results on the three reverberated conditions. The "no front-end" configuration accuracies are very similar across the three T_{60}s, and significantly lower than the "Clean" condition ones. In the FMD task, the accuracy decreases by 31.37% on average while in the FLD task by 34.41%. A similar performance drop can be observed in the majority agreement tasks. This behaviour is due to both the reverberation effect, and to the presence of all the participants' voices in each input signal, which makes it impossible for the dominance detector to discriminates the four voices. The introduction of the speech enhancement front-end significantly improves the detection results, giving an accuracy improvement of 32.68% in the FMD task and of 36.56% in

Table 2. Dominance detection results. See Sec. 4.1 for the task labels description.

Accuracy (%)	Task	120 ms	240 ms	360 ms	Average
No front-end	FMD	50.00	55.88	55.88	53.92
	FLD	48.39	45.16	45.16	46.24
	MMD	53.57	55.36	51.79	53.57
	MLD	42.59	37.04	35.19	38.27
Full-system	FMD	82.35	85.29	91.18	86.61
	FLD	83.87	80.65	83.87	82.80
	MMD	76.79	76.79	82.14	78.57
	MLD	64.81	62.96	66.67	64.81

the FLD one. In the majority agreement results, the behaviour is similar: in the MMD task the improvement is 25.00%, while in the MLD task is 26.54%.

Note, finally, that both in the "Clean" and reverberated conditions, the majority agreement results are lower on average than the full agreement ones. This is due to the higher variability in the annotations and is consistent with [3].

5 Conclusion

This work presented a dominance detection framework able to operate in multi-talker reverberated acoustic scenarios. The overall framework is composed of two main blocks, a speech enhancement front-end and a dominance detector. The task of the first is to reduce the reverberation effect induced by the convolution between the meeting participant voice signals and the room impulse responses. This is performed using a recently proposed solution [7] composed of three stages: speaker diarization, room impulse response identification and speech dereverberation. The dominance detection algorithm is based on the speech feature extraction toolkit openSMILE. To exploit contextual information, a bidirectional Long Short-Term Memory network which produces the final estimate of the activity level for each speaker is employed. Experiments have been performed on the DOME dataset: results obtained on reverberated versions of the corpus have shown the effectiveness of the developed system, making it appealing for applications in real-life human-computer interaction scenarios.

Future developments will involve both the dominance estimator and the speech enhancement front-end. The first will be augmented with video features, which have been already successfully exploited in the literature [2,4]. The feature set could be also augmented with the speaking lengths of each participant coming from the speaker diarizer. In addition, the evaluation of the so-called bottleneck network architectures for enhanced BLSTM modelling of a participant's activity in meetings is planned. With regard to the front-end, the presence of additive noise will be considered and suitable procedures will be taken into account to reduce its impact. Moreover, the speaker diarization stage will be featured with an overlap-detector, which also allows to include a source separation stage within the front-end and exploit also the overlapped speech segments.

References

1. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: A review. Image and Vision Computing 27(12), 1775–1787 (2009)
2. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations using nonverbal activity cues. IEEE Trans. on Audio, Speech, and Language Processing 17(3), 501–513 (2009)
3. Hung, H., Huang, Y., Friedland, G., Gatica-Perez, D.: Estimating dominance in multi-party meetings using speaker diarization. IEEE Trans. on Audio, Speech, and Language Processing 19(4), 847–860 (2011)
4. Aran, O., Gatica-Perez, D.: Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In: Proc. of Int. Conf. on Pattern Recognition, pp. 3687–3690 (August 2010)
5. Aran, O., Hung, H., Gatica-Perez, D.: A multimodal corpus for studying dominance in small group conversations. In: LREC Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, Malta (May 2010)
6. Hörnler, B., Rigoll, G.: Multi-modal activity and dominance detection in smart meeting rooms. In: Proc. of ICASSP, pp. 1777–1780 (2009)
7. Rotili, R., Principi, E., Squartini, S., Schuller, B.: A Real-Time Speech Enhancement Framework for Multi-party Meetings. In: Travieso-González, C.M., Alonso-Hernández, J.B. (eds.) NOLISP 2011. LNCS, vol. 7015, pp. 80–87. Springer, Heidelberg (2011)
8. Huang, Y., Benesty, J.: A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Speech Audio Process. 51(1), 11–24 (2003)
9. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE - the Munich versatile and fast open-source audio feature extractor. In: Proc. of ACM Multimedia, Firenze, Italy, pp. 1459–1462 (2010)
10. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: AVEC 2011–The First International Audio/Visual Emotion Challenge. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 415–424. Springer, Heidelberg (2011)
11. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE Journal of Selected Topics in Signal Processing 4(5), 867–881 (2010)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997)
13. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18(5-6), 602–610 (2005)
14. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI Meeting Corpus: A Pre-announcement. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 28–39. Springer, Heidelberg (2006)
15. Johnson, D.H., Dudgeon, D.E.: Array Signal Processing. Prentice-Hall, Englewood Cliffs (1993)