Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200 (B.C.))

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST – Priority 2

DELIVERABLE NO: D1.8
Incongruence Detection for Detecting, Removing, and
Repairing Incorrect Functionality in Low-Level Processing

Date of deliverable: 31.12.2009
Actual submission date: 08.02.2010

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable:
Czech Technical University (CTU)

Revision [0]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Incongruence Detection for Detecting, Removing, and Repairing Incorrect Functionality in Low-Level Processing

Tomáš Pajdla[1], Michal Havlena[1], Jan Heller[1],
Hendrik Kayser[2], Jörg-Hendrik Bach[2], Jörn Anemüller[2]

December 26, 2009

Abstract

The theory of incongruence, which deals with inconsistent decisions of direct and composite classifiers of the same concept, can be used to improve low-level processing by detecting incorrect functionality and repairing it through re-defining the composite classifier. In this report, we summarize the advancements in the direct audio and the direct audio-visual classifiers yielding two speaker detectors which can cause an incongruence. Then, we show how this incongruence could be used for learning a new concept, direct position consistency classifier, which can be used to re-define the composite speaker classifier.

1 Audio-Visual Speaker Detector

The audio-visual speaker detector presented last year was further updated in order to handle multiple speakers and a moving observer. First, we recapitulate the relationship of the audio-visual speaker detector to the theory of incongruence introduced in [12].

1.1 Application of the Theory of Incongruence

Figure 1 shows an example of the "Speaker event" that is recognized in two ways, either by a direct classifier, which is trained directly from complete audio-visual data, or by a composite classifier that evaluates the conjunction of "Human sound event" and "Human look event" direct classifiers.

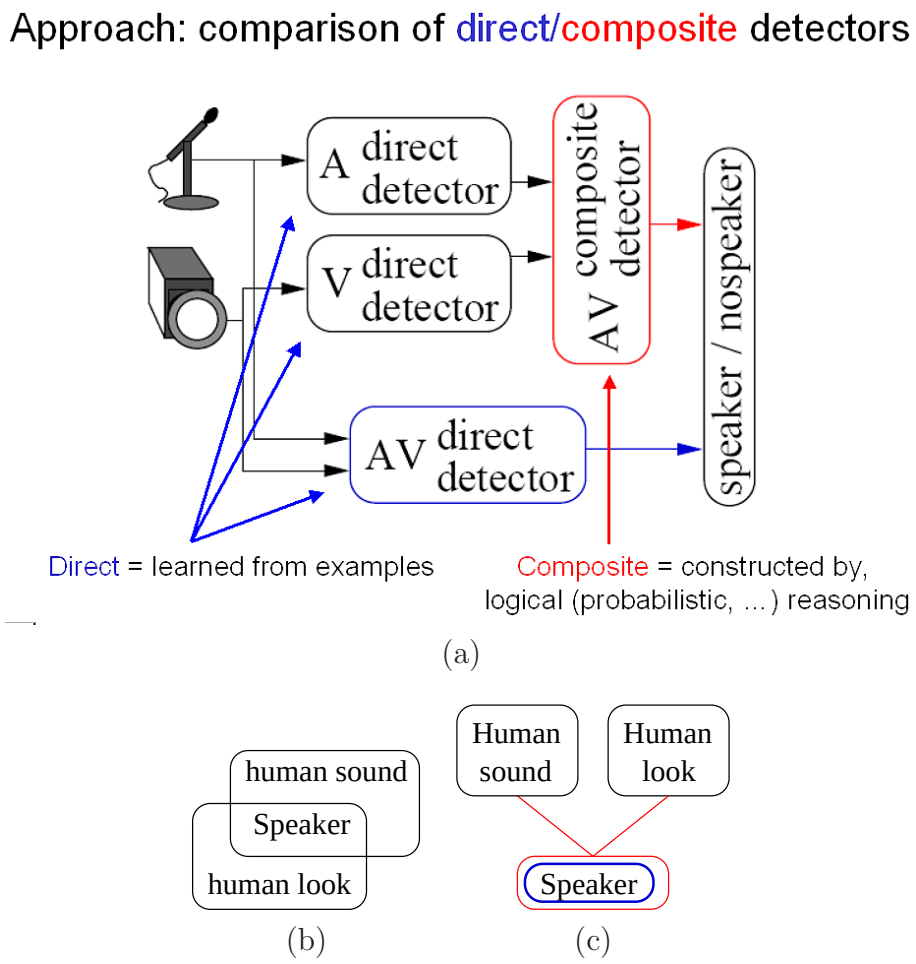


Figure 1: (a) "Speaker" is recognized in two ways, either by a holistic (direct) classifier, which is trained directly from complete audio-visual data, or by a composite classifier, which evaluates the conjunction of "Human sound" and "Human look" direct classifiers. (b) "Speaker" is given by the intersection of sets representing "Human sound" and "Human look", which corresponds to the infimum in the Boolean POSET (c).

"Speaker" is given by the intersection of sets representing "human sound" and "human look" which corresponds to the infimum in the Boolean POSET. In the language of [13, 18], the composite classifier corresponds to the general level (i.e. to $Q^g_{speaker}$) while the direct classifier corresponds to the specific level (i.e. to $Q_{speaker}$).

The direct audio classifier (see Section 1.2, Figure 5) detects human sound, e.g. speech, and returns a boolean decision on the "Human sound

		Composite Q_a^g (general level)	Direct Q_a (specific level)	Possible reason
1	$0 \simeq Q_a^g \simeq Q_a \simeq 0$	reject	reject	empty silent scene
2	$1 \simeq Q_a^g \gg Q_a \simeq 0$	accept	reject	silent person speaking loudspeaker
3	$0 \simeq Q_a^g \ll Q_a \simeq 1$	reject	accept	inconsistent POSET wrong model
4	$1 \simeq Q_a^g \simeq Q_a \simeq 1$	accept	accept	speaking person

Table 1: Interpretation of agreement/disagreement for the "Speaker event" example.

event". The direct visual classifier (see Section 1.3, Figure 11) detects human body shape in an image and returns a boolean decision on the "Human look event". The direct audio-visual classifier (see Section 1.4, Figures 3 and 15) detects the presence of a speaker and returns a boolean decision on the "Speaker event". The composite audio-visual classifier (see Section 1.5, Figures 4 and 16) constructed as the conjunction of the direct audio and visual classifiers also detects the presence of a speaker and returns a boolean decision on the "Speaker event". Opposed to the direct audio-visual classifier, its decisions are constructed from the decisions of the separate classifiers using logical AND.

After presenting a scene with a silent person and speaking loudspeaker, the composite audio-visual classifier fires but the direct audio-visual classifier does not give a positive answer. That creates a disagreement, incongruence, between classifiers.

Table 1 interprets the results of speaker detection.

1.2 Direct Audio Detector

The direct audio classifier for sound source localization is based on features extracted from the generalized cross correlation (GCC) function [9] between two audio input signals. A sound source is localized by estimating the azimuthal angle of the direction of arrival (DOA) relative to the sensor plane defined by the two front-microphones of the AWEAR 2.0 platform.

Support-vector-machines (SVM) [1] are used to classify the presence or absence of a source at each angle. This approach enables the simultaneous localization of more than one sound source in each time-frame.

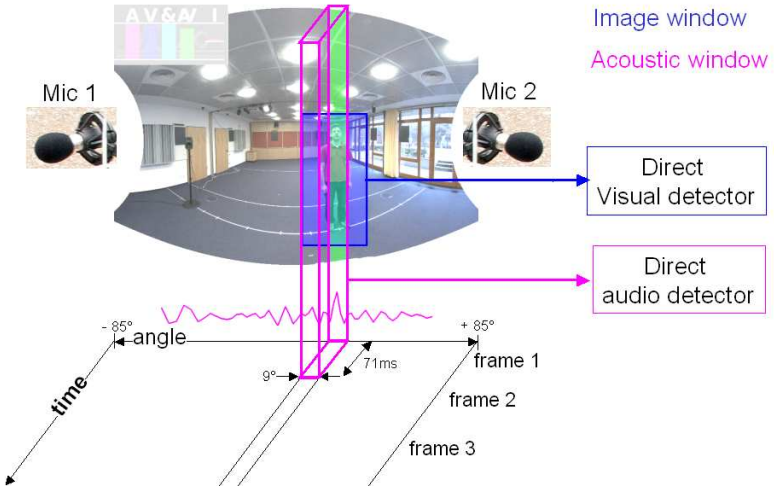


Figure 2: Detectors are constructed to work on windows that are scanned across the field of view. This approach is motivated by successful approach to face and human body detection.

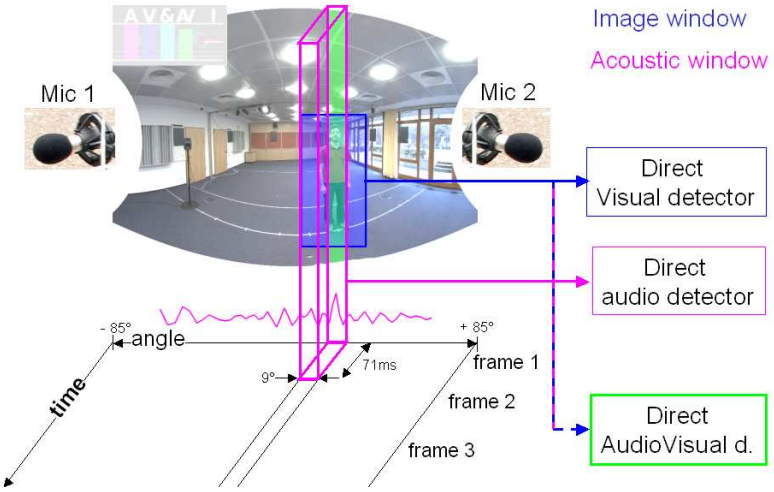


Figure 3: The direct audio-visual detector takes features extracted from the audio and video signals to train a discriminative classifier separating events concurrent in space and time from other events. This detector captures a compressed form of the information relevant to audio-video detections but can't really explain them in terms of simpler modules, i.e. outputs of separated direct audio and video detectors.

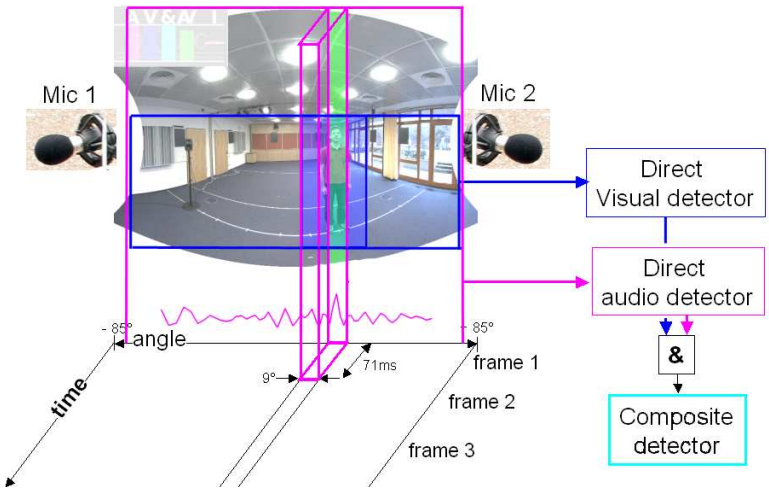


Figure 4: The composite audio-visual detector is the conjunction of the direct audio and video detectors, each evaluated per complete field of view. Notice that the composite detector detects events concurrent in time but not necessarily co-located in the field of view. It will fire on sound coming from a loudspeaker on the left and a silent person standing on the right when that happens in the same moment.

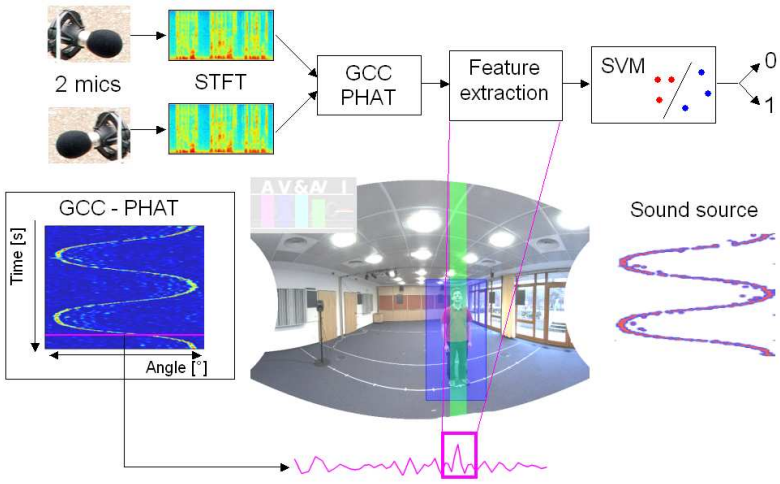


Figure 5: The direct audio detector as described in the text.

1.2.1 Localization Feature Extraction

The GCC is an extension of the cross power spectral density function, which is given by the Fourier transform of the cross correlation. Given two signals

$x_1(n)$ and $x_2(n)$, it is defined as:

$$G(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H_1(\omega) H_2^*(\omega) \cdot X_1(\omega) X_2^*(\omega) e^{j\omega n} \, \mathrm{d}\omega, \tag{1}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the respective signals and the term $H_1(\omega)H_2^*(\omega)$ denotes a general frequency weighting.

In the present work PHAse Transform (PHAT) weighting [9] has been used, which normalizes the amplitudes of the input signals to unity in each frequency band, $H_1(\omega)H_2^*(\omega) = \frac{1}{\left|X_1(\omega)X_2^*(\omega)\right|}$:

$$G_{PHAT}(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(\omega) X_2^*(\omega)}{|X_1(\omega) X_2^*(\omega)|} e^{j\omega n} \, \mathrm{d}\omega, \tag{2}$$

such that only the phase difference between the input signals is preserved.

The audio data is captured with a sampling frequency of 48 kHz. The inter-microphone distance in the AWEAR setup (45 cm) corresponds to a maximum delay of 1.32 ms or 64 samples ($\widehat{=}\ \pm 90°$) in each direction or 128 samples in total. The window length and the length of the Fourier transform used to compute the GCC in the spectral domain are chosen to be 512 samples, the overlap between consecutive windows is 50%. This yields a 257-dimensional GCC-PHAT vector for every time frame of 5.33 ms length.

A frequency cut-off of 8 kHz is applied prior to the GCC calculation. Final spectral-domain zero-padding results in a time-domain resolution for the delay estimates of twice the original sampling rate.

From the GCC data, angular features are extracted for classification. The feature extraction algorithm is as follows. A 16 samples long rectangular window is slid over the GCC-vector with a shift of 4 samples, subdividing the data of a single time frame into 61 feature vectors. This covers the field of view homogeneously in terms of time delay with a resolution of $\frac{1}{3}$ ms. The mapping from the time delay τ to the angle of incidence θ is non-linear: $\theta = \arcsin(\tau \cdot \frac{c}{d})$, where c denotes the speed of sound and d the distance between the sensors. This results in a non-homogenous angular resolution in the DOA-angle-space, with higher resolution near the center and lower resolution towards the edges of the field of view (see Figure 6). If at least one of these 61 directional feature vectors is classified positively, the direct audio classifier fires.

1.2.2 Source Localization through Classification

To produce a suitably large set of training data, sound sources were simulated by using mono-channel speech recordings from the TIMIT speech corpus [6]

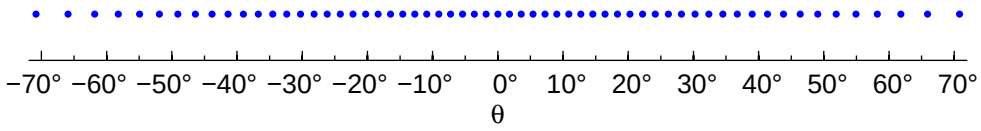


Figure 6: Distribution of the angles of incidence for the DOA classification.

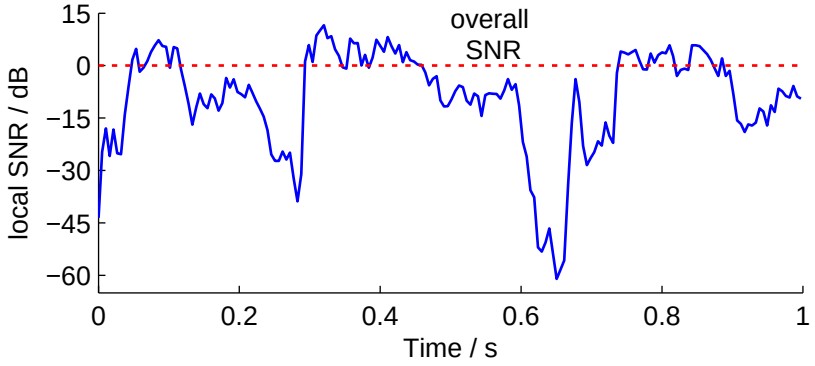


Figure 7: Local SNR for a speech signal in a diffuse noise field with an long-term SNR of 0 dB.

and generating a second channel as a delayed version of the same data, thus introducing directional information into the data. The simulated DOA angle ranged from $-80°$ to $+80°$ in steps of $10°$, for each direction 10 seconds of speech were used. A spherically symmetric diffuse pink noise field was generated and superposed on the speech signals with varying long-term SNR ranging from -20 dB to +20 dB in steps of 5 dB.

To account for local drops of the SNR within speech pauses, a short-time SNR in each time frame was introduced, referred to as local SNR (see Figure 7). Subsequently, the training material was labeled as 'source is present' if the delay between left and right audio channel corresponded to the respective angle *and* the local SNR exceeded a fixed threshold. This threshold was varied from -20 dB to +20 dB in steps of 5 dB. For each long-term SNR and each local-SNR threshold one model was trained on all available training directions, resulting in 81 models for all conditions.

1.2.3 Selection of the Local-SNR Threshold

In order to evaluate which local-SNR threshold yields the best model performance in terms of classification and generalization, the models were tested on a different set of simulated data. The test data were also taken from the

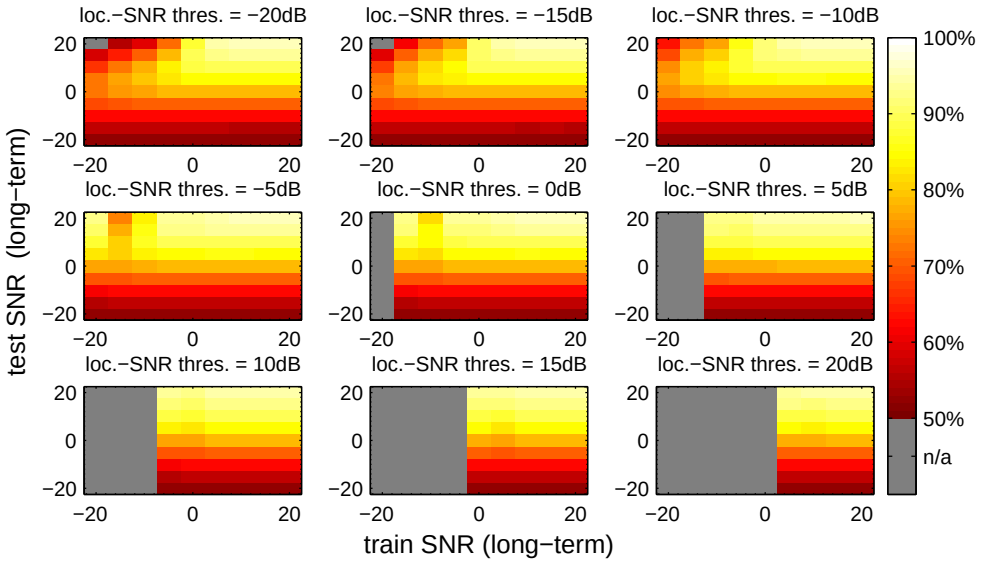


Figure 8: Results for all models tested on all SNRs for each local-SNR-threshold during training, given in percentage of correct decisions of the classifier.

TIMIT speech corpus. The DOA angle contained in the test data ranged from $-80°$ to $+80°$ in steps of $5°$. The same long-term SNRs as before were used. For testing, however, the data were labeled without using a local-SNR criterion.

To evaluate the robustness against noise and the generalization performance of the models, each model (trained with a fixed local-SNR threshold) was tested on all available SNRs. The results are given in the percentage of correct decisions, see Figure 8.

These results indicate that above a certain local-SNR threshold the performance of the classification is nearly independent of the training SNR. -5 dB turns out to be a reasonable choice for the threshold. If the local-SNR threshold is set too high, training of the models became impossible for low SNRs due to the resulting lack of training data, as in this case to much data are rejected as feature for a present source.

A classification example for two speech sources with interference by a noise field at an SNR of 5 dB is shown in Figure 9. The noisy GCC-matrix is given at the left and the results of the classification at the right. The theoretical position of the two speech sources at $-20°$ and $+45°$ are indicated by the dotted lines in the right panel. To suppress the "salt and pepper" noise contained in the results of the classification, median filtering over 3 adjacent angles and 7 time frames was applied.

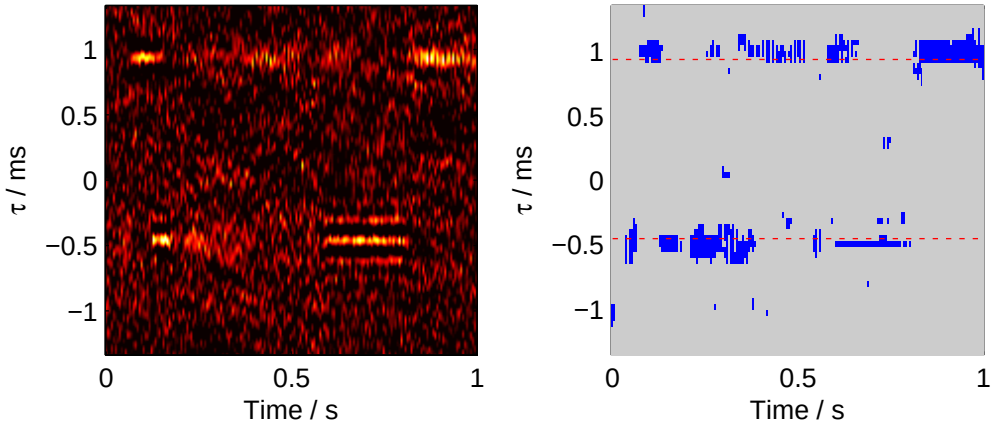


Figure 9: GCC matrix (lefthand) and results of the classification (righthand) for two speech sources at $-20°$ and $+45°$ in a isotropic noise field at an SNR of 5 dB. The dotted lines in the right plot denote the theoretical positions of the sources.

1.2.4 Speech Detection

The localization algorithm described above reliably detects several sound sources at different positions. It does not, however, include a characterization of said sources. In order to provide an acoustic "human detector", a second step is necessary to classify the type of sound originating from the localized source. If the detected class is speech, the output of the direct audio classifier contains the information that a person was detected at a certain position.

The features used in this task are Amplitude Modulation Spectrograms (AMS, see Deliverable 2.11), which are motivated by the importance of modulations identified in numerous psychophysical, physiological and applied studies [4, 8, 10, 16]. AMS represent a decomposition of the signal along the dimensions of acoustic frequency, modulation frequency and time, and are computed by a (modulation) spectral decomposition of sub-band spectral power time-courses in overlapping temporal windows.

The processing stages of the AMS computation are as follows (see Figure 10). The signal decomposition with respect to acoustic frequency is computed by a short-term fast Fourier transformation (FFT with 32 ms Hann window, 4 ms shift, FFT length 256 samples (32 ms), sampling rate 8 kHz.), followed by squared magnitude computation, summation into rectangular, non-overlapping Bark bands and logarithmic amplitude compression.

Within each spectral band, the modulation spectrum is obtained by applying another FFT (1000 ms Hann window, 500 ms shift, FFT length 250 samples (1000 ms)) to the temporal trajectories of subband log-energy. Out-

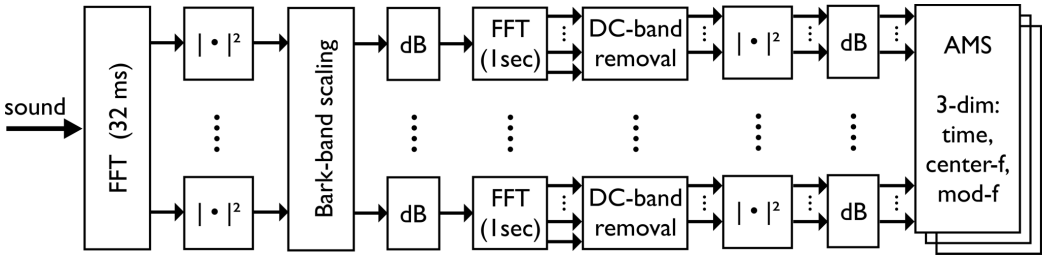


Figure 10: The AMS feature extraction algorithm.

puts in the 0 Hz and 1 Hz modulation bands are influenced by DC components in the (log-energy) spectral domain and discarded as a means to reduce effects of channel noise (see also below). Finally, an envelope extraction and a further logarithmic compression are applied. The resolution of 17 Bark acoustic frequency bands and 29 modulation frequency bands (2 Hz to 30 Hz) results in a signal representation with 493 feature values per time step, at a temporal resolution of 1000 ms.

Robustness of the AMS features with respect to constant and slowly-varying time-domain convolutive factors (channel noise) is enhanced by construction of the feature extraction scheme: The first Fourier transform step approximately converts the convolution to a multiplicative term in each (acoustic) frequency band. The subsequent logscale compression converts it into an additive term that is captured in the DC (0 Hz) modulation spectral band after the second FFT (depending on parameter choice, additional low modulation-frequency may be influenced by smearing of DC components). Deliberately discarding the affected modulation bands therefore results in AMS features that are approximately invariant to a time-domain signal convolution with short impulse responses such as microphone transfer functions and early reverberation effects.

Using these features, an SVM is trained on the two classes "speech in background" and "background sound only". The training data consists of a set of recordings obtained within DIRAC as background sounds (road noise and pedestrian zone noise) and the TIMIT data base as speech source. The training SNR has been optimized in a parameter search over a range of +20 dB to -20 dB in steps of 5 dB. -5 dB turns out to yield the best overall performance on various test data.

This model is used as speech detector in the DIRAC recordings, indoor as well as outdoor. The robustness against variation in the SNR and variation in the type of acoustic background has been systematically evaluated (see Deliverable 2.11).

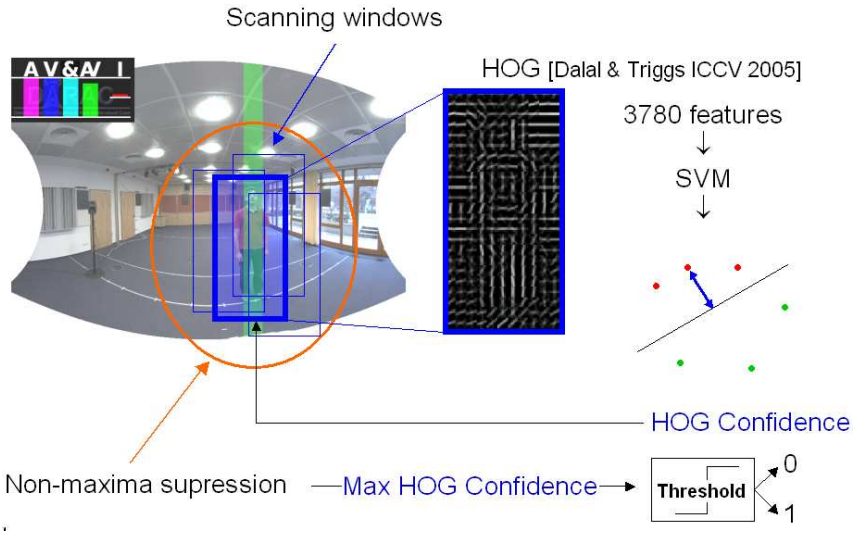


Figure 11: The direct visual detector is based on work [3].

1.3 Direct Visual Detector

The direct visual classifier remained the same as the one presented in [12] as it performed quite well already. The state-of-the-art paradigm to visual human detection [3] classifies every feasible window in the image for presence or absence of a human-like shape by adopting the assumption that the ground plane is parallel to the image, which is true for our static camera scenario. Therefore, one can restrict feasible rectangles to those that correspond to reasonably tall (150–190 cm) and wide (50–80 cm) pedestrians standing on the ground plane. HoG (see Figure 12) visual features can be computed in each such rectangle as described below.

The INRIA OLT detector toolkit [11] based on the histograms of oriented gradients (HoG) algorithm presented by Dalal and Triggs in [3] chosen to detect humans uses a dense grid superimposed over a detection window to produce a 3780 dimensional vector to train a linear SVM classifier. A detection window of 64×128 pixels is divided into cells of 8×8 pixels, each group of 2×2 cells then integrated into overlapping blocks. Each cell consists of 9-bin HoG that are concatenated for each block and normalized using L^2 norm. Each detection window consists of 7×15 blocks resulting in 3780 features. See Figure 12 for an example of the feature vector of a detection window. As the cylindrical projection images used in our experiment are locally similar to the perspective ones, the detector was trained on perspective images.

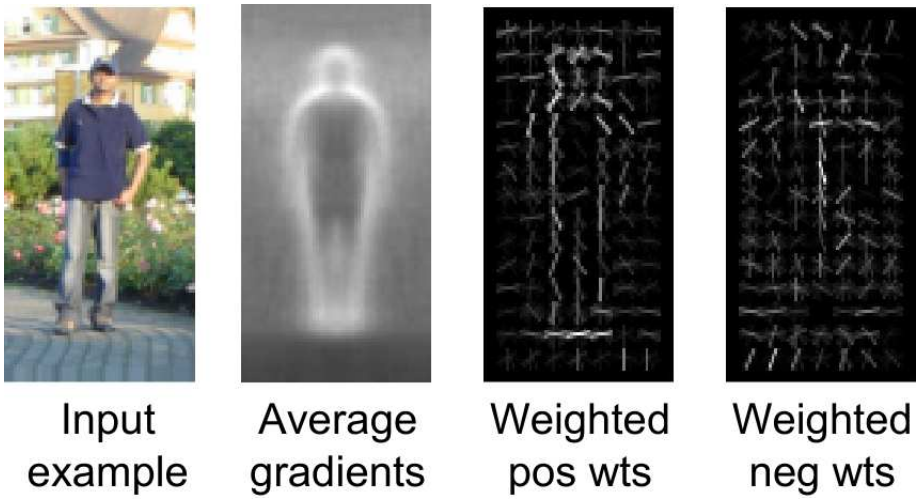


Figure 12: Histograms of oriented Gradients (HoG) features form the base of the direct visual classifier. Image courtesy of [3].

The tested image is scanned by the detection window at several scales and window positions with detection scores (confidences) higher than 0, i.e. tentative human detections, are further processed by non-maximum suppression using mean shift robust mode detection into final human detections. If the detection score of the best resulting human detection is higher than 0.1, the direct visual classifier fires.

1.4 Direct Audio-Visual Detector

For the direct audio-visual classifier, we use the concept of angular (azimuthal) bins that allows for handling multiple pedestrians and/or sound sources which was not possible with the classifier presented last year. 180° field of view is divided into twenty bins, 9° each, and the classification is performed per bin, see Figure 13. If at least one of the bins is classified positively, the direct audio-visual classifier fires. The procedure of the classification of each bin follows.

First, a 2D feature vector is constructed from audio and visual features as the highest GCC-PHAT value and the highest pedestrian detection score belonging to the bin. The pedestrian detection score is maximized both in the x and y coordinates of the window center where x has to lie in the bin and y goes through the whole height of the image as the bins are azimuthal (i.e. not bounded in the vertical direction). Then, the feature vector is classified by an SVM classifier with the RBF kernel [15]. Any non-negative SVM score yields a positive classification.

Figure 13: 180° horizontal field of view is divided into twenty bins, 9° each. Each bin is described by a 2D feature vector constructed from audio and visual features as the highest GCC-PHAT value and the highest pedestrian detection score belonging to the bin. Notice different widths of the individual bins explained in Section 1.6.1

The SVM classifier was trained from four sequences (2,541 frames in total) of people speaking while walking along a line there and back. The positive bin was labeled manually for each frame and the 2D feature vectors belonging to the bin were added to the positive examples. Two neighboring bins from each side were excluded and the feature vectors belonging to the rest of the bins were added to the negative examples. As we needed also examples of people not speaking and sounds without people which were not in the training data, we created negative examples for these cases by combining parts of the positive and negative feature vectors yielding 109,023 negative examples in total. The training data together with the trained decision curve can be seen in Figure 14.

1.5 Composite Audio-Visual Detector

The composite audio-visual detector is constructed to explain audio-visual events by a combination of simpler audio and visual events. The direct audio-visual detector primarily separates relevant events from the irrelevant

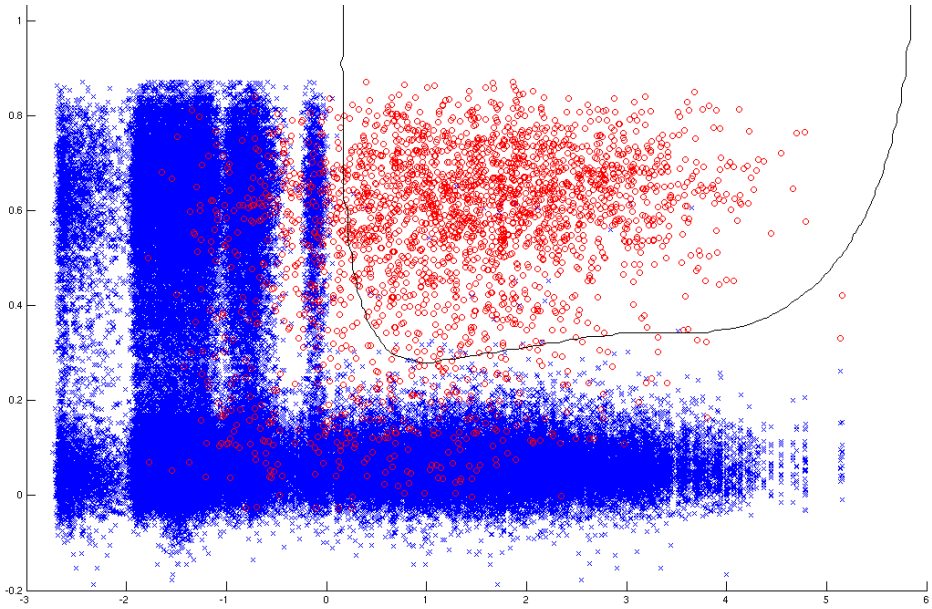


Figure 14: Training data and the trained decision curve for SVM with the RBF kernel of the direct audio-visual classifier. GCC-PHAT values (x-axis) and pedestrian detection scores (y-axis) for different positive (red circles) and negative (blue crosses) manually labeled examples.

ones. That is achieved by presenting the relevant events in the positive subset of the training set while the other (irrelevant) ones in the negative one. The composite audio-visual detector, on the other hand, represents "the understanding to the world".

Currently, the two direct audio and video detectors are combined using two-state logic, which is too simplistic to cope with full complexity of real situations but it provides all basic elements of reasoning that later could be modeled probabilistically.

The composite detector, Figures 4 and 16, is not perfect. It is constructed as the conjunction of the direct audio and video detectors, each evaluated per complete field of view. Hence, it does not capture spatial co-location of the audio-visual events. It detects events concurrent in time but not necessarily co-located in the field of view, Figure 16. It will fire on sound coming from a loudspeaker on the left and a silent person standing on the right when that happens at the same moment.

This insufficiency is intentional. It models a current understanding of the world, which is due to its complexity always only partial, and leads to detecting incongruence w.r.t. the direct audio-visual detector when human sound and look come from dislocated places.

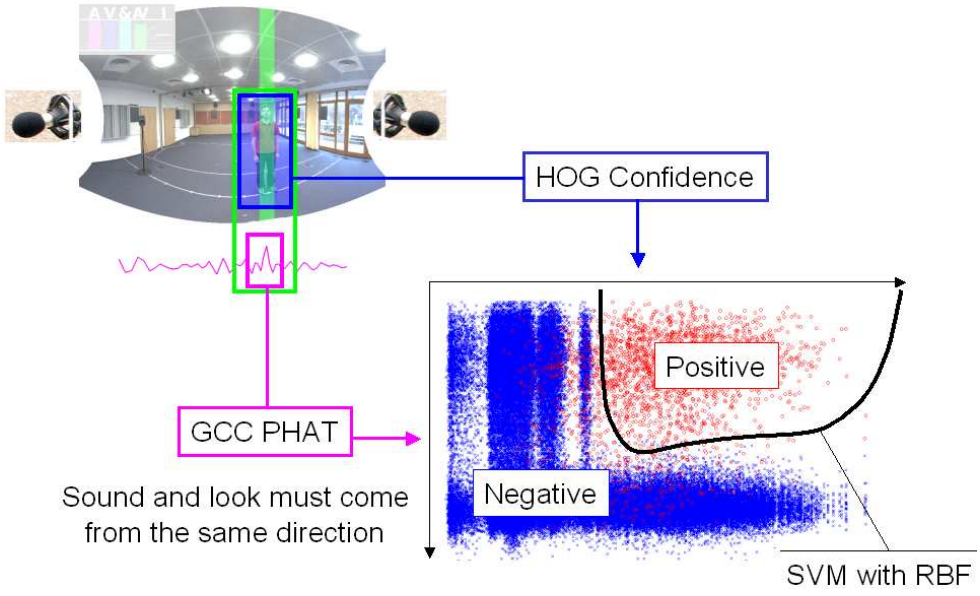


Figure 15: The decisions of the direct audio-visual detector.

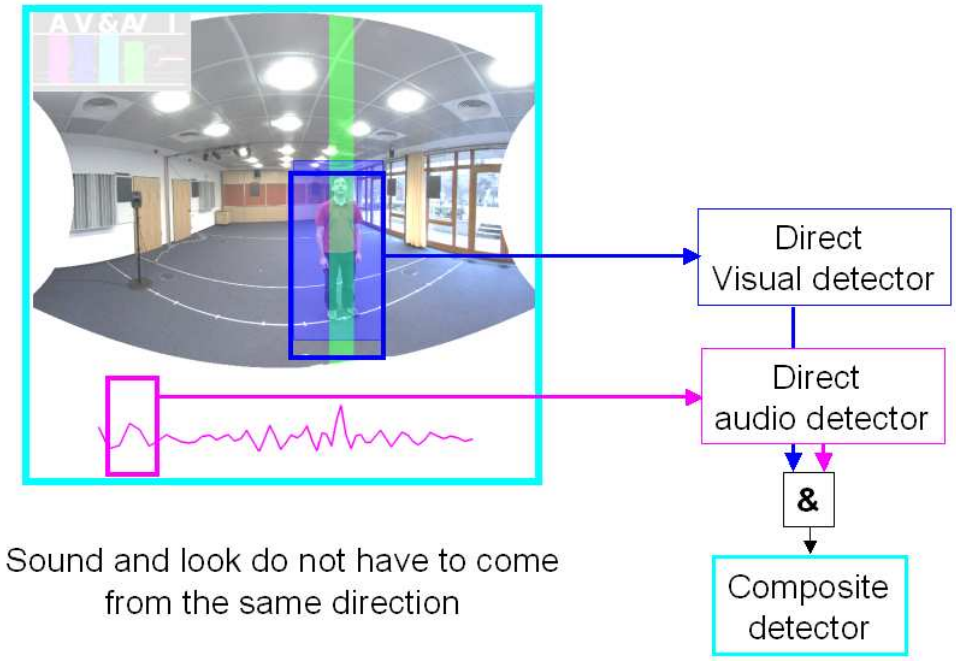


Figure 16: The decisions of the composite audio-visual detector.

(a) (b)

Figure 17: Two example frames from SPEAKER&LOUDSPEAKER sequence showing a (a) congruent and an (b) incongruent situation when the speaker and the loudspeaker sound respectively. Decisions from different direct classifiers are drawn into the frames: direct audio classifier (magenta), direct visual classifier (blue), direct audio-visual classifier (green). The bars in the top-left corner show also the composite audio-visual classifier decisions (cyan) and incongruence/wrong model (red/yellow).

1.6 Experimental Results

The direct and composite audio-visual classifiers were used to process several sequences acquired by the AWEAR 2.0 platform [7]. They contain a person speaking while walking freely in the view-field. After a while, the person stops talking and the loudspeaker sounds up, rendering an incongruent event which should be detected. The description of the extension for a moving observer comes after the experiment with the static device.

1.6.1 Static Device

As the video data come from the left camera of a stereo rig with a 45 cm wide baseline, there is a discrepancy between the camera position and the apparent position of a virtual listener to which the GCC-PHAT is computed, which is the center of the acquisition platform. To compensate for this error, the distance to the sound source and the distance between the virtual listener and the camera need to be known. The listener–camera distance can be computed as 22.5 cm from the known setup of the rig. The distance to the sound source is now assumed to be 1.5 m from the camera. The corrected

(a) (b) (c)

Figure 18: Pedestrian detection in frames stabilized w.r.t. the ground-plane transferred back to the original frame. (a) Non-stabilized frame. (b) Stabilized frame with pedestrian detection. (c) Non-stabilized frame with the transferred pedestrian detection. Notice the non-rectangular shape of the pedestrian detection after the transfer.

angle can be then trivially computed from the camera–listener–sound source triangle using a line–circle intersection.

The accuracy of the incongruence detection is lower for speakers much further away than the assumed 1.5 m due to the aforementioned angle correction, nevertheless, the detector performed quite well in our experiments. Two example frames from a 30 sec long sequence SPEAKER&LOUDSPEAKER shot at 14 fps can be found in Figure 17.

1.6.2 Moving Observer

The script of the 73 sec long 14 fps sequence MOVINGOBSERVER is similar to the one of SPEAKER&LOUDSPEAKER sequence, the difference being the fact that the AWEAR 2.0 device was worn (i.e. moving and slightly rocking) when acquiring it. As the microphones are rigidly connected with the cameras, the relative audio-visual configuration remains the same and movement does not cause any problems to the combination of the classifiers. On the other hand, the direct visual classifier is able to detect upright pedestrians only, therefore we perform pedestrian detection in frames stabilized w.r.t. the ground-plane [17] and transfer the results back into the original frames yielding non-rectangular pedestrian detections, see Figure 18.

As the x coordinate of the mouth can be different than the x coordinate of the center of the detected window for non-upright pedestrians, we use the position of the center of the upper third of the window (i.e. the estimated mouth position) in the direct audio-visual classifier instead, see Figure 19 for two example frames from the sequence.

(a) (b)

Figure 19: Two example frames from MOVINGOBSERVER sequence showing (a) a congruent and (b) an incongruent situation when the speaker and the loudspeaker sound respectively. Notice the non-rectangular shape of the pedestrian detections. See Figure 17 for the color legend.

2 Learning from Incongruence

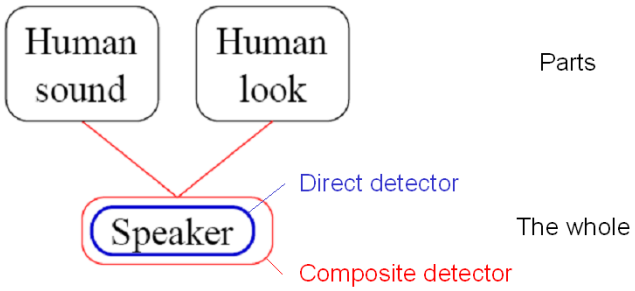
Figure 20 illustrates a prototypical system consisting of alternative detectors, which can lead to a disagreement between the alternative outcomes related to an event.

Three direct detectors and one composite detector is shown in Figure 20. The direct detector of "Human sound", the direct detector of "Human look", the direct detector of "Speaker" and the composite detector of "Speaker" are presented. The composite detector, as explained above in more detail, was constructed as a logical combination of direct detectors evaluated on the whole field of view, hence not capturing the spatial co-location of sound and look events defining a speaker in the scene.

The table in Figure 20 shows the four possible combinations of outcomes of the direct and composite "Speaker" detectors as analyzed in [18].

The first row, Figure 21(a), where none of the detectors fire, corresponds to no event, noise or completely new concept, which has not been yet learned by the system. The last row, Figure 21(a) again, when both detectors fire, corresponds to detecting a known concept.

The second row, when the "Speaker" composite detector fires but the direct one remains negative, corresponds to the *incongruence*. This case can be interpreted as having a partial model of a concept, e.g. not capturing some important aspect like the spatial co-location. Alternatively, it also can



Weinshall et al. NIPS 2008

	Direct (specific)	Composite (general)	Possible reason
1	reject	reject	New concept a, noisy X
2	reject	accept	Incongruence
3	accept	reject	Wrong model
4	accept	accept	Known concept

Figure 20: Alternative detectors of the "Speaker" concept and their possible outcomes. See text.

happen when the model of the concept is wrong such that it mistakingly requires some property which is not truly related to the concept.

The third row, Figure 21(b), when the direct "Speaker" detector fires but the composite one remains negative, corresponds to the *wrong model* case. Indeed, this case applies when the composite detector mistakingly requires some property which is not truly related to the concept. However, it happens also when the composite detector has only a partial model of the concept, e.g. when it misses one of possible cases in which the concept should be detected.

We can see that the interpretation of the second and third rows depends on how the composite detectors are constructed. Restricting ourselves to Horn clauses [2], which is a popular choice since it allows efficient derivation and is used in PROLOG [14], will choose the interpretation and will make then the distinction between the second and third rows to correspond to the interpretation from [18].

Assume to have a composite C detector constructed in the form of a Horn clause of direct detectors $D_1, D_2, \ldots, D_n$

$$D_1 \wedge D_2 \wedge \ldots D_n \rightarrow C \tag{3}$$

which means that C is active if and only if all D_i are active. For instance "Human look" $\wedge$ "Human sound " $\rightarrow$ "Speaker" is a Horn clause.

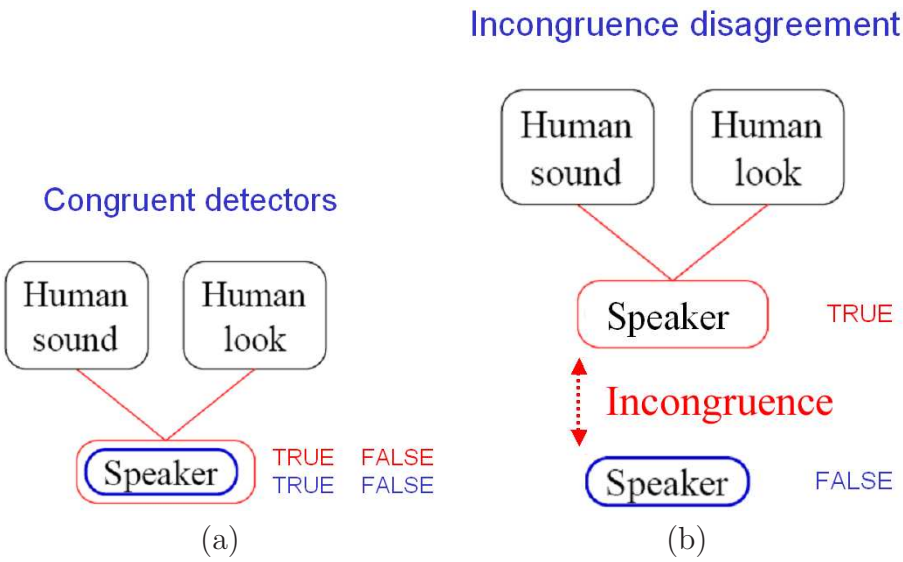


Figure 21: (a) Congruent result of detectors. (b) Incongruence is detected when the direct detector rejects more (i.e. is smaller) than the composite detector.

With this restriction, a detected incongruence can be understood as if the composite detector missed a term on the left hand side of the conjunction in the derivation rule, which is responsible for rejecting the falsely accepted cases. It is easy to remedy this situation by learning a new concept corresponding to the missing term in the conjunction from the wrongly classified examples.

There are many possibilities how to do it. A particularly simple way would be to add a single new concept "Co-located" to the conjunction, Figure 22, i.e.

$$\text{“}Human\ look''\wedge\text{“}Human\ sound''\wedge\text{“}Co{-}located''\rightarrow\text{“}Speaker'' \qquad (4)$$

which would "push" the composite detector "down" to coincide with the direct detector.

Somewhat more redundant but still feasible alternative would be to add two more elements to the system as shown in Figure 23. A new composite detector "H. S. & L.", i.e. "Human Sound and Look" could be established and combined with another newly introduced concept "Co-located" to update the model in order to correspond to the evidence. Although somewhat less efficient, this second approach may be preferable since it keeps concepts for which detectors have been established already.

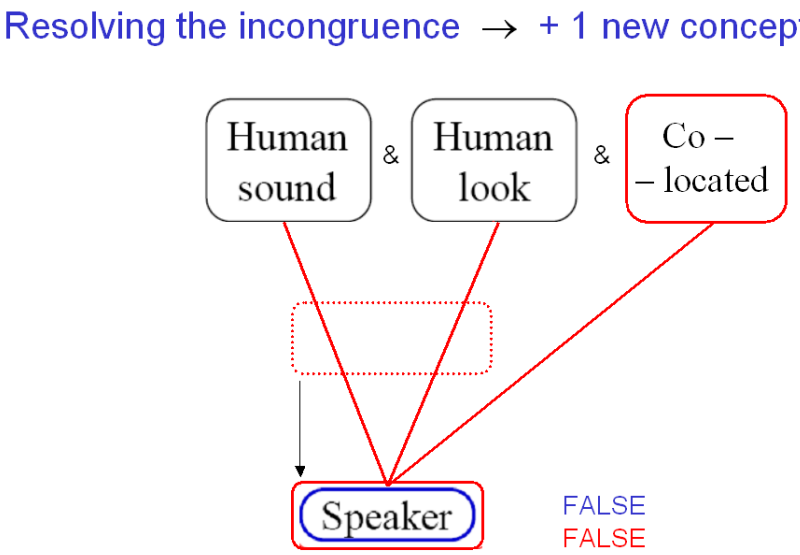


Figure 22: Adding a single new concept "Co-located" to the conjunction defining the "Speaker" concept.

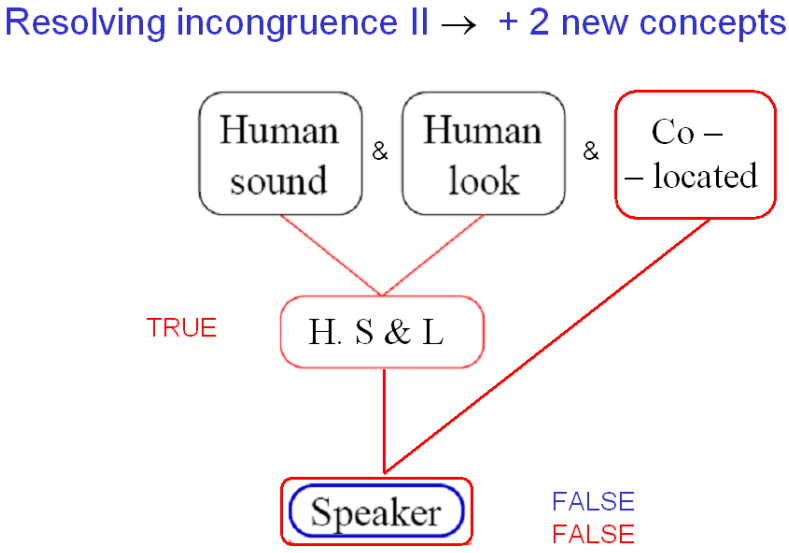


Figure 23: H. S. & L.–"Human Sound and Look" concept can explain the conjunction of existing detectors. The "Speaker" detector can be then constructed by adding the new "Co-located" concept.

As suggested above, the incongruence, i.e. the disagreement between the direct and the composite classifiers, may signal that the composite classifier is not well defined. We would like to use the incongruent data to learn a new concept, which could be used to re-define the composite classifier and to remove the incongruence.

In the case of the speaker detector, the composite audio-visual classifier has to be re-defined. A new "Human Sound and Look" concept has to be initiated. The composite audio-visual classifier is disassociated from the "Speaker" concept, new "Human Sound and Look" concept is created and associated to the composite classifier. This new concept will be greater than the "Speaker" concept. Next, a new composite audio-visual classifier is created as a conjunction of the old composite audio-visual classifier and a classifier deciding a new "XYZ" concept which needs to be trained using the incongruent data. The new composite classifier is associated to the "Speaker" concept. The name of the "XYZ" concept can be established later based on its interpretation by a human.

2.1 Learning the Direct Position Consistency Classifier

We will deal with the simplest case when the incongruence is caused by a single reason which can be modeled as a new concept. Our goal is to establish a suitable feature space and to train a direct classifier deciding the "XYZ" concept using the congruent and incongruent data as positive and negative training examples respectively. As the values of audio and visual features have been used in direct audio and direct visual classifiers already and our new concept should be as general as possible, we will use only boolean values encoding the presence of a given event in the 20 angular bins.

First, two feature vectors of length 20 are created for each frame, one encoding the presence of audio events and the other one encoding the presence of visual events, and concatenated together in order to form a boolean feature vector of length 40.

Secondly, in order to find dependencies between the different positions of the events, the feature vector is lifted to dimension 820 by computing all possible products between the 40 values. The original feature vector:

$$x_1\ x_2\ x_3\ \ldots\ x_{39}\ x_{40}$$

is transformed into:

$$x_1^2\ x_1x_2\ x_1x_3\ \ldots\ x_2^2\ x_2x_3\ \ldots\ x_{39}x_{40}\ x_{40}^2.$$

When the original values are boolean, the quadratic monomials $x_1^2, x_2^2, \ldots$ have values equal to $x_1, x_2, \ldots$ and the monomials $x_1x_2, x_1x_3, \ldots$ have values equal to the conjunctions $x_1 \wedge x_2, x_1 \wedge x_3, \ldots$. SVM training over these vectors should reveal significant pairs of positions by assigning high weights to the corresponding positions in the lifted vector.

We use two sequences to construct the positive and negative training example sets. Each of these sequences is nearly 5 minutes long with a person walking along the line there and back with a loudspeaker placed near one of the ends of the line. During the first approx. 90 seconds, the walking person is speaking and the loudspeaker is silent rendering a congruent situation. During the next approx. 90 seconds, the walking person is silent and the loudspeaker is speaking causing an incongruent situation. In the last approx. 90 seconds, both the walking person and the loudspeaker are speaking which is congruent by our definition as we are able to find a bin with a speaker.

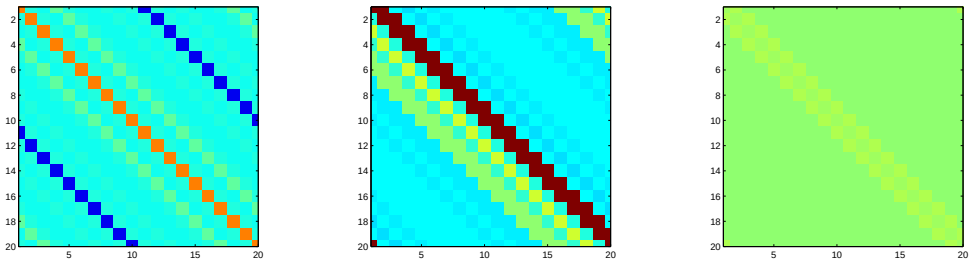
For each frame, the concatenated boolean feature vector is created. The boolean decisions for the 61 directional feature vectors from the audio detector are transformed into the 20 values of the audio part of the feature vector using disjunction when more directional labels fall into the same bin. The visual part of the feature vector is initialized with 20 zeros and each confident human detection output by the visual detector changes the value belonging to the angular position of the center of the corresponding rectangle to one. Feature vectors belonging to frames which yielded a positive response from both the composite and the direct audio-visual classifiers (i.e. congruent situation) are put into the positive set, those belonging to frames that were classified positively by the composite but negatively by the direct audio-visual classifier (i.e. incongruent situation) are put into the negative set, and those belonging to frames with a negative response from the composite audio-visual classifier are discarded as such data cannot be used for our training.

As the loudspeaker position is fixed in our training sequences, we decided to remove the bias introduced by this fact by "rotating" the data around the bins, so each training example is used to generate 19 other training examples before lifting, e.g. a training example:

$$x_1\ x_2\ x_3\ \ldots\ x_{20}\ x_{21}\ \ldots\ x_{39}\ x_{40}$$

is used to generate 19 additional examples:

$$\begin{gathered} x_2\ x_3\ x_4\ \ldots\ x_1\ x_{22}\ \ldots\ x_{40}\ x_{21} \\ x_3\ x_4\ x_5\ \ldots\ x_2\ x_{23}\ \ldots\ x_{21}\ x_{22} \\ \vdots \\ x_{20}\ x_1\ x_2\ \ldots\ x_{19}\ x_{40}\ \ldots\ x_{38}\ x_{39}. \end{gathered}$$



(a) (b) (c)

Figure 24: Resulting weights for different pairs of values in the feature vector as obtained by SVM. Numbers on axes correspond to bin indices. Positive weights are denoted by red, zero weights by green and negative weights by blue color. (a) Audio × audio. (b) Audio × visual. (c) Visual × visual.

Finally, the feature vectors of 60,320 positive and 41,820 negative examples were lifted and used to train a linear SVM classifier [5].

The results shown in Figure 24 can be commented as follows. The most significant result is the dark red main diagonal in the A×V diagram (Figure 24(b)) telling us that the positive examples have the audio and visual events in the same bin (or shifted by one bin as one of the neighboring diagonals is red too). Red square at $(1, 20)$ is a by-product of the "rotation" as neighboring bins can be separated to different ends of the view-field.

As can be seen in the V×V diagram (Figure 24(c)), pairs of visual events are insignificant. The orange main diagonal in the A×A diagram (Figure 24(a)) says that positive examples tend to contain more audio events. This is due to the fact that the only the situation with two audio events present in the training data was congruent, we had no training data with two loudspeakers speaking. The light blue adjacent diagonal is also an artifact of the direct audio-visual detector and "rotation".

To conclude, the just trained classifier decides the position consistency of the audio and visual events, so a suitable name for the "XYZ" concept would be the "Co-located" concept.

3 Conclusion

In this report, we have made a step toward resolving incongruence arising from a disagreement between direct and composite classifiers. We have identified that the interpretation of the disagreement between the direct and composite classifiers can be interpreted in alternative ways depending on the

way how the composite classifiers are constructed but we have also seen that the interpretation accepted in DIRAC is the one corresponding to adopting Horn clauses for constructing the logical model of observations.

Next, we have carried out an experiment showing how to cope with an incongruence and demonstrated that it can be removed from the system by adding new concepts and new detectors.

There are many interesting open questions such as whether the Horn clauses are practical enough to model interesting situations in our applications, how to select which concepts should be represented and which should not, or how to best construct new detectors when lifting becomes expensive.

References

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] Horn Clause. *Wikipedia*. `http://en.wikipedia.org/wiki/Horn_clause`.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*, volume 2, pages 886–893, 2005.

[4] G. Evangelopoulos and P. Maragos. Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. Aud. Sp. Lang*, 14(6):2024–2038, 2006.

[5] V. Franc and S. Sonneburg. Optimized cutting plane algorithm for large-scale risk minimization. *Journal of Machine Learning Research*, 10:2157–2232, October 2009.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. U.S. Dept. of Commerce NTIS, Gaithersburg, MD, 1990.

[7] M. Havlena, A. Torii, W. Moreau, and T. Pajdla. Omnidirectional audio-visual data acquisition and processing. Research Report CTU–CMP–2008–26, CMP Prague, December 2008.

[8] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel. On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, 28:43–55, 1999.

[9] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, August 1976.

[10] B. Kollmeier and R. Koch. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *Journal of the Acoustical Society of America*, 95:1593–1602, 1994.

[11] Inria OLT. `http://pascal.inrialpes.fr/soft/olt/`, 2008.

[12] T. Pajdla, L. Van Gool, M. Havlena, J. Heller, A. Torii, A. Ess, J.-H. Bach, H. Kayser, J. Anemüller, and P. Van Hengel. Incongruence detection in audio-visual processing. Research Report CTU–CMP–2008–28, CMP Prague, December 2008.

[13] M. Pavel, H. Jimison, D. Weinshall, A. Zweig, F. Ohl, and H. Hermansky. Detection and identification of rare incongruent events in cognitive and engineering systems. Dirac white paper, OHSU, April 2008.

[14] PROLOG. *Wikipedia*. `http://en.wikipedia.org/wiki/Prolog`.

[15] B. Schölkopf and A. Smola. *Learning with Kernels*. The MIT Press, MA, 2002.

[16] J. Tchorz and B. Kollmeier. SNR estimation based on amplitude modulation analysis with applications to noise suppression. *IEEE Transactions on Speech and Audio Processing*, 11(3):184–192, 2003.

[17] A. Torii, M. Havlena, and T. Pajdla. Omnidirectional image stabilization by computing camera trajectory. In *PSIVT 2009*, pages 71–82, 2009.

[18] D. Weinshall et al. Beyond novelty detection: Incongruent events, when general an specific classifiers disagree. In *NIPS 2008*, pages 1745–1752, 2008.