Speech Emotion Recognition using Affective Saliency

Arodami Chorianopoulou[1], *Polychronis Koutsakis*[4], *Alexandros Potamianos*[2,3]

[1] School of ECE, Technical University of Crete, Chania 73100, Greece
[2]School of ECE, National Tecnical University of Athens, Zografou 15780, Greece
[3]Athena Research and Innovation Center Maroussi 15125, Athens, Greece
[4]School of Engineering and Information Technology, Murdoch University, Murdoch, Australia, 6150

achorianopoulou@isc.tuc.gr, p.koutsakis@murdoch.edu.au, apotam@central.ntua.gr

Abstract

We investigate an affective saliency approach for speech emotion recognition of spoken dialogue utterances that estimates the amount of emotional information over time. The proposed saliency approach uses a regression model that combines features extracted from the acoustic signal and the posteriors of a segment-level classifier to obtain frame or segment-level ratings. The affective saliency model is trained using a minimum classification error (MCE) criterion that learns the weights by optimizing an objective loss function related to the classification error rate of the emotion recognition system. Affective saliency scores are then used to weight the contribution of frame-level posteriors and/or features to the speech emotion classification decision. The algorithm is evaluated for the task of anger detection on four call-center datasets for two languages, Greek and English, with good results.

Index Terms: affective saliency, emotion recognition, fusion over time, spoken dialogue systems

1. Introduction

Research by psychologists and neuroscientists has shown that emotion is an important aspect of human interaction, as it is highly related to decision-making. In Spoken Dialogue Systems (SDS) the analysis of speakers' emotion [1, 2, 3], age, gender [4] or personality [5] can significantly improve dialogue management strategies and improve the user experience. Affective systems perform acoustic and linguistic analysis to assign a variety of categorical labels to emotional states or estimate continuous emotional scores.

Identifying signal features suitable to describe affective information is challenging. The standard approach in emotion recognition systems is to extract prosodic features, particularly pitch and energy [6, 7, 8]. In [9] Mel-Frequency Cepstral coefficients (MFCCs) have been used for training acoustic and phonetic tokens, while in [10] contextual features were proposed for spoken dialogue systems, including prosodic and discourse context.

Several machine learning techniques have been also explored for affective modeling. Support Vector Machines (SVM) [11], Hidden Markov Models (HMMs) [12], and Gaussian Mixture Models (GMMs) [13] are proposed for speech emotion recognition. In [14] the emotion recognition performance was compared using SVM, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) classifiers, while segment level approaches are also introduced to model the emotional aspects of the speech signal in [15]. In other paralinguistic tasks, e.g., cognitive load estimation, i-vectors have also been investigated [16].

One of the main issues in affective classification is the level (phone, utterance) of information integration and decision fusion, as well as how information over different time-scales is fused over time. The most popular information fusion method for affective computing is feature-level fusion, where statistics of frame-level features (low-level descriptors) are estimated over a segment or for the whole utterance. In [17], a number of fusion methods are presented, while in [18] decision fusion over different modalities is presented. Applying a discriminative procedure as Minimum Classification Error (MCE) training [19, 20] for information fusion over time has been investigated in the past for several tasks including automatic speech recognition and speaker recognition [21]. In [22] spectral distance features combined with a frame-level misclassification error have been investigated for information fusion over time using conditional random field classifiers. Such techniques are shown to reduce the classification error rate significantly and increase the discriminability among the different labels.

In this work, we present a model for information fusion over time that weights speech frames/segments based on their affective saliency. This fusion is implemented following either an early (feature-level) or a late fusion scheme. Affective saliency is estimated via a regression model that utilized features extracted from different timescales of the acoustic signal (e.g., F0) and the frame-level posterior probabilities. The regression model is trained using a Minimum Classification Error (MCE) criterion. The method iteratively updates the trainable parameters, in order to minimize the classification error rate. In our experiments, we used spoken dialogue call-center datasets and we focus on an anger detection task (negative vs non-negative valence detection).

The remainder of the paper is organized as follows. The proposed system is presented in Section 2. The saliency model, classification and information fusion shemes are then analyzed in Section 3. The datasets and experimental procedure are shown in Section 4. Finally, results are presented in Section 5, while conclusions are provided in Section 6.

2. System Description

The system's main components are presented in Figure 1. First a frame-level feature vector is constructed. It is assumed that each frame contains an expression of the emotion of the utterance it belongs to, and therefore it is given that same label. The resulting feature vector with the assumed frame-level labels is then given as input to train a frame-level classifier. The frame-level decisions of a given utterance are further combined in a weighting scheme, which emphasizes the most salient affective information over time. This weighting scheme is trained via a regression model with features derived from the frame-level acoustic features. The regression parameters are trained iteratively by minimizing the classification error rate via MCE training/ Generalized Probabilistic Descent (GPD) [23]. The utterance-level emotion decision is then computed according to two scenarios, an early (feature-level) or a late fusion scheme.

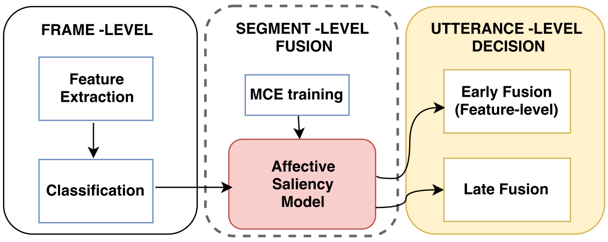

Figure 1: System architecture

3. Affective Saliency Model

Let $X = \{x_1, \ldots, x_N\}$ be a frame vector of an utterance T, and C_i discrete affective labels, e.g. levels of anger vs. neutral, with $i = 1, \ldots, M$. The emotional content of an utterance T is computed over time by its corresponding frames and weighted according to the factor λ_j which indicates the affective saliency for frame j.

$$F(C_i|X) = \log P(C_i|X) = \frac{1}{N}\sum_{j=1}^{N} \lambda_j \log P(C_i|x_j) \quad (1)$$

where $P(C_i|x_j)$ are the frame-level posterior probabilities, while the weights λ_j are estimated via Minimum Classification Error (MCE). More specifically, given that the optimal weights are unknown, we train a regression model as:

$$\lambda_j = \sum_{k=1}^{K} a_k d_k \quad (2)$$

where a_k with $\sum_{k=1}^{K} a_k = 1$ the trainable weights and d_k the regression features, described in Section 4.1.2. The next step is to define the misclassification measure E, as shown below

$$E(X) = F(C_I|X) - F(C_C|X) \quad (3)$$

where C_I and C_C correspond to the incorrect and correct emotional classes, respectively. The loss function, which maps the misclassification error onto the interval $[0, 1]$ is a sigmoid function and it is defined as

$$l(X) = \frac{1}{1 + e^{-\gamma E(X)}}, \qquad \gamma > 1 \quad (4)$$

with γ representing the sigmoid scaling factor. The loss function approaches zero when $E(X) < 0$ and close to one otherwise. So by minimizing the loss function, the classification error is also minimized. The loss function $l(X)$ can be differentiated and optimized via an iterative gradient descent algorithm, by establishing the algorithmic convergence property [23]. The update equation of a specific unknown parameter w is

$$w' = w - \epsilon \frac{1}{N_T} \sum_{\forall T} \frac{\partial l(X)}{\partial w} \quad (5)$$

where N_T is the total number of utterances T in the dataset, ϵ is a learning rate parameter used during the iterative MCE training and $\frac{\partial l(X)}{\partial w}$ the partial derivative of the loss function $l(X)$

$$\frac{\partial l(X)}{\partial w} = \frac{\partial l(X)}{\partial E(X)} \cdot \frac{\partial E(X)}{\partial \lambda_j} \cdot \frac{\partial \lambda_j}{\partial w} \quad (6)$$

3.1. Late Fusion

First we investigate a late fusion scheme for the utterance-level emotion decision. Specifically, we combine the computed weights λ_j as shown in Eq. (2) with the frame-level posterior probabilities of our affective classifier $P(C_i|x_j)$, as presented in Eq. (1). Then the utterance-level emotion decision is computed as:

$$C^* = \arg\max_{C_i} F(C_i|X) \quad (7)$$

where C_i, with $i = 1, \ldots, M$ the discrete affective labels.

3.2. Early Fusion (Feature-level)

The saliency weights are used to compute weighted statistics over the frames of an utterance, namely mean, standard deviation, max, min and median. Given a frame $j, 1 \leq j \leq N$, with feature value f_j and weight w_j the weighted mean $\hat{\mu}_w$ and standard deviation σ_w are:

$$\mu_w = \frac{\sum_{j=1}^{N} w_j f_j}{\sum_{j=1}^{N} w_j}, \qquad \sigma_w = \sqrt{\frac{\sum_{j=1}^{N} w_j (f_j - \mu_w)^2}{\sum_{j=1}^{N} w_j}} \quad (8)$$

The weighted median is estimated as feature values f_j that can appear multiple times, according to their weights w_j.

4. Experimental Procedure

4.1. Affective Saliency Experiments

Initially, features have been normalized in the [0,1] interval across all the utterances of a dataset both for the affective and the regression model.

4.1.1. Affective Saliency Classification

For the affective classification defined in (1), we found that the trainable parameters were more robust across datasets when computed on segment-level instead of frame-level. Hence, features were grouped in sets of 20 frames and statistics were computed over them. We use only 3 LLDs, namely energy, 1st Mel-Frequency Cepstral Coefficient (MFCC) and raw fundamental frequency (F0) and applied the following statistics: max, min, mean, median and standard deviation.

4.1.2. Regression Features

In this section we present the parameter estimation model and the saliency features d_k, as described in Eq. (2). Several features including features derived from the posterior probabilities and the acoustic signal were also evaluated as candidates for estimating affective saliency. We found that spectral flux and F0 extracted from different timescales of the speech signal, were robust across the different datasets. Specifically, we extracted spectral flux and F0 in a fixed window size of 200 ms and F0 in 30 ms with 10 ms update. Features extracted in 30 ms window size were further grouped in order to create segments and statistics were applied, namely max, min, mean, median, standard deviation. As an additional feature, we used the rate of unvoiced frames per segment using the Voice Activity Detector presented in [26].

4.1.3. Optimization and Parameter Estimation

During MCE-training the a_k parameters were iteratively updated. In each iteration the average loss value was shown to decrease while the classification accuracy increased, as more misclassified utterances were corrected. The optimal parameters are the ones that minimized the average loss function. The scaling factor γ of Eq. (4) and learning factor ϵ of Eq. (5) were set to $\gamma = 2$ and $\epsilon = 0.1$. We observed that for both matched and cross experiments (see Section 4), after 300 iteration the GPD algorithm converges for the selected parameters γ and ϵ.

The parameters a_k were initially trained independently on each dataset to investigate the robustness of the proposed method. Results were pretty consistent across dataset. Finally we selected the median value across the datasets in order to construct a universal saliency model. The resulting weights for the $[0, 1]$ normalized features are presented in Table 1.

F0 (30ms)					200ms		
max	min	med.	std	mean	Spec. Flux	F0	Unv. Rate
0.21	0.09	0.20	0.04	0.17	0.21	0.09	0.11

Table 1: Estimated optimal parameters across all datasets for the matched experiments.

Figure 2 shows the speech signal and the frame-level pitch contour of the utterance “No, can I talk to a person?” with the weights λ_j computed according to Eq. (2). The weights are computed on segment-level and mapped to samples and/or frames using linear interpolation. The weights' values vary across time and peaks are detected toward the end of the utterance where the word ”person” is stressed (see also F0 contour). The saliency curve is very smooth since the saliency weights are computed on segment-level.

4.2. Affective Feature Extraction

A set of 33 frame-level features (low-level descriptors) and their deltas were extracted in a fixed window size of 30 ms with a 10 ms frame update, using the OpenSmile toolkit. The list of spectral and prosodic features used is given in Table 2.

Energy-related LLDs	Energy, Zero-Crossing Rate
Spectral LLDs	Energy 250-650Hz 1k-4kHz, Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity, MFCC 1-14, Roll Off Point 0.25, 0.50, 0.75, 0.90
Voicing realted LLDs	F0, Prob. of Voice, raw F0

Table 2: List of features

Regarding the baseline and early fusion scenarios the features in Table 2 were used along with their deltas. Similar to the saliency model (described in Section 4.1), features have been mapped into the [0,1] interval. In order to extract utterance-level features, the following functionals were applied: mean, standard deviation, median, max and min.

4.3. Data

For our experiments we used four spoken dialogue datasets from four call-centers in two languages: (1) bus information (LEGO, a subset of the Let'sGo dataset [24]), (2) US call center (CC) incoming customer service calls, (3) phone banking (PB) [25] and (4) movie ticketing (MT) [25]. CC was annotated in a binary scale: angry vs neutral. LEGO, PB and MT datasets were annotated using a 5-level scale for anger detection: *friendly, neutral, slightly angry, angry, very angry*. These labels were then mapped to two classes; *friendly, neutral* mapped to the non-negative class and *slightly angry, angry, very angry* to the negative. A brief description of the datasets is presented in Table 3.

	LEGO	CC	PB	MT
#non-negative	3309	1027	1095	1023
#negative	934	339	607	1106
#speakers	200	284	[1]	200
Language	English	English	Greek	Greek

Table 3: Dataset description.

4.4. Experiments

We conducted two types of experiments across all datasets: matched (training and testing on the same corpus) and cross-corpus. In the matched experiments, we divided each dataset in equally sized training, development and test sets, while for the cross-corpus experiments, we used (all the data of) three datasets for training and development and tested on the fourth. The development set was used for learning the unknown parameters a_k of Eq. (2). Table 4 presents the average utterance duration per dataset, which as expected is an important factor for the model's performance.

	CC	LEGO	PB	MT
Average duration	1.85	1.67	4.17	1.43

Table 4: Average utterance duration in seconds per dataset.

Regarding the experimental procedure, the chance classifier assigns each test sample to the majority class. For our baseline experiments as well as the feature-level fusion an SVM classifier with polynomial kernel from the Weka toolkit is used [27]. We chose an SVM classifier due to its better performance compared to other classifiers tested. Additionally, a forward selection algorithm from the Weka toolkit was applied on the baseline system and the selected features were adapted on the early fusion scenario as well. For the saliency model we chose a Naive Bayes classifier, in order to extract the class-posterior probabilities, and we present results before (pre-MCE) and after (post-MCE) MCE training.

5. Evaluation & Results

Next, we present the unweighted average *(UA)* classification accuracy across all datasets and fusion scenarios for the matched and cross-corpus experiments.

In Table 5 the results for the late fusion scenario are presented for both the matched and cross experiments. The re-

[1] No information about the number of speakers was available for the phone banking dataset.

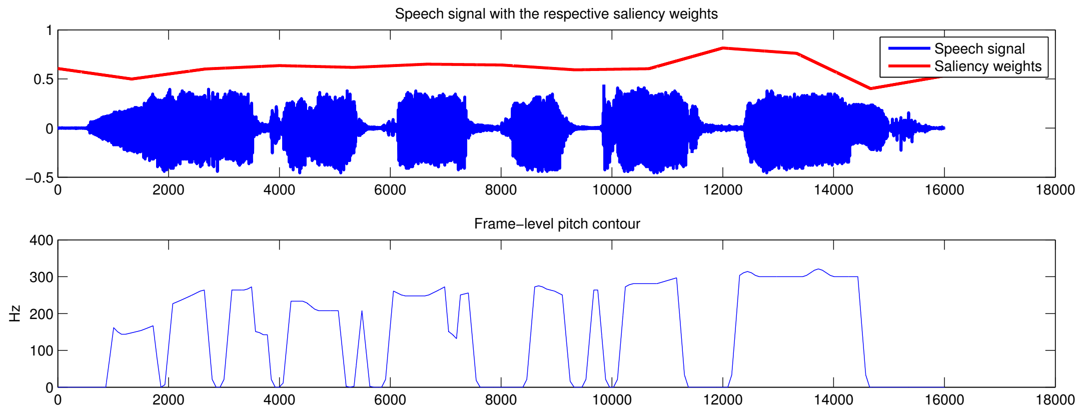

Figure 2: Utterance of the CC dataset with transcription: "No, can I talk to a person?". Estimated affective saliency (top) and fundamental frequency contour (bottom) is also shown.

	CC	LEGO	PB	MT	UA
Matched experiments					
pre-MCE	77.4	78.7	68.8	53.4	69.5
post-MCE	80.5	79.6	68.1	52.7	70.2
Cross-corpus experiments					
pre-MCE	81.4	79.0	65.6	58.0	71.0
post-MCE	81.6	79.5	66.0	58.2	71.4

Table 5: Late fusion: Classification accuracy (%) results for the matched and cross experiments.

gression model (affective saliency weights) is initially trained independently by minimizing the average loss function on each dataset and further estimated across all datasets. Results are presented before (no weighting) and after MCE training. As we can see the MCE approach has better performance than the pre-MCE system when refering to the UA metric. When comparing each dataset's performance individually, for the cross-corpus post-MCE outperforms pre-MCE for all experiments, although the improvement is small.

	CC	LEGO	PB	MT	UA
Chance	73.4	79.4	64.2	52.7	67.4
Baseline	79.2	79.8	67.6	51.7	69.6
Early fusion	80.0	80.3	68.2	51.7	70.1

Table 6: Early fusion: Classification accuracy (%) results for the matched experiments.

	CC	LEGO	PB	MT	UA
Chance	75.2	77.9	64.3	51.9	67.3
Baseline	81.6	82.1	66.3	54.0	71.0
Early fusion	80.8	82.5	66.7	57.8	72.0

Table 7: Early fusion: Classification accuracy (%) results for the cross-corpus experiments.

In Table 6 the results of the early (feature-level) fusion are presented for the matched experiments. For both the baseline and the fusion system, statistics are applied to frame-level LLDs in order to extract utterance-level features. However, for the feature-level fusion weighted statistics are used. The weights are computed according to the saliency model and mapped to frame-level using linear interpolation. We observe equal or better performance for each dataset individually, suggesting that the global nature of the affective saliency system is robust across the different datasets.

Table 7 shows the classification accuracy results for the early fusion scenario on the cross-corpus experiments. Here the affective model is computed on three datasets and tested on a fourth. We observe similar behavior with the results presented in Table 6, which suggests robustness across the different datasets. This is impressive given that our datasets are of different languages, sizes and SDS type.

Overall, we show improvement across all datasets using the affective saliency model either with the early or the late fusion fusion scenarios, suggesting that frame-level decisions can be fused more efficiently in order to characterize the utterance-level emotional content.

6. Conclusions

We investigated the automatic recognition of emotions in speech using an affective saliency model for fusing information over time. The proposed fusion algorithm exploits an affective saliency regression model to either weight frame-level posterior classification probabilities or frame-level features. We demonstrated that the proposed model can achieve modest performance improvement over the baseline. Our results suggest that MCE training increases the discriminability between emotional states, by enhancing the speech frames that carry the most salient information. In future work, a richer feature set and alternative machine learning algorithms will be evaluated for affective fusion.

7. Acknowledgements

This work has been partially supported by the SpeDial project supported by the EU FP7 with grant no. 611396 and the Baby-Robot project supported by the EU Horizon 2020 Programme with grant no. 687831.

8. References

[1] Busso, C., Bulut, M., and Narayanan, S., “Toward effective automatic recognition systems of emotion in speech”, Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds., pp. 110-127, 2012.

[2] Galanis, D., Karabetsos, S., Koutsombogera, M., Papageorgiou, H., Esposito, A., and Riviello, M., “Classification of Emotional Speech Units in Call Centre Interactions”, IEEE 4th International Conference on Cognitive Infocommunications, pp. 403-406, 2013.

[3] Kim, S., Georgiou, P. G., Lee, S., and Narayanan, S., “Real-time Emotion Detection System using Speech: Multi-modal Fusion of Different Timescale Features”, IEEE 9th Workshop on Multimedia Signal Processing, 2007.

[4] Meinedo, H. and Trancoso, I., “Age and gender classification using fusion of acoustic and prosodic features”, in Proc. INTERSPEECH, pp. 2818-2821, 2010.

[5] Mohammadi, G., and Vincianelli, A., “Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features”, IEEE Transactions on Affective Computing, 3(3), pp. 273-284, 2012.

[6] Ververidis, D., Kotropoulos, K., and Pittas, I., “Automatic Emotional Speech Classification”, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, pp. 593-596. 2004.

[7] Lee, C. M., Narayanan, S., and Pieraccini, R., “Recognition of Negative Emotions from the Speech Signal”, IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 240-243, 2001.

[8] Schuller, B., Lang, M., and Rigoll, G., “Automatic Emotion Recognition by the Speech Signal”, Institute for Human-Machine-Communication, Technical University of Munich, 2002.

[9] Li, M., “Automatic Recognition of Speaker Physical Load using Posterior Probability Based Features from Acoustic and Phonetic Tokens”, in Proc. INTERSPEECH, pp. 437-441, 2014.

[10] Liscombe, J., Riccardi, G., and Hakkani-Tr, D, “Using context to improve emotion detection in spoken dialogue systems”, in Proc. INTERSPEECH, pp. 18451848, 2005.

[11] Lee, C. M., Yildirim, S., Bulut, M., and Narayanan, S., “Emotion recognition based on phoneme classes”, in 8th International Conference on Spoken Language Processing, pp. 889-892, 2004.

[12] Nwe, T., Foo, S., and De Silva, L., “Speech emotion recognition using hidden Markov models”, Speech Communications, 41(4), 603623, 2003.

[13] Busso, C., Lee, S., and Narayanan, S., “Analysis of emotionally salient aspects of fundamental frequency for emotion detection”, IEEE Transactions on Audio, Speech and Language Processing, 17(4), pp. 582596, 2009

[14] Kwon O., Chan K., Hao J., and Lee T., “Emotion recognition by speech signals”, in Proc. of Eurospeech, pp. 125-128, 2003.

[15] Shami, M., and Verhelst, W., “An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech”, Speech Communication 49, pp. 201-212, 2007.

[16] Van Segbroeck, M., Travadi, R., Vaz, C., Kim, J., Black, M. P., Potamianos, A. and Narayanan, S. S., “Classification of Cognitive Load from Speech using an i-vector Framework”, in Proc. of InterSpeech, 2014.

[17] Ruta, D., and Gabrys, B., “An overview of classifier fusion methods”, Computing and Information systems, pp. 1–10, 2000.

[18] Metallinou, A., Lee, S. and Narayanan, S., “Decision level combination of multiple modalities for recognition and analysis of emotional expression”, in Proc. of ICASSP, pp. 2462–2465, 2010.

[19] Juang, B. H., and Katagiri, S., “Discriminative Learning for Minimum Error Classification”, IEEE Transactions on Signal Processing, 40(12), 1992.

[20] Ephraim, Y., Dembo, A. and Rabiner, L. R., “A minimum discrimination information approach for hidden Markov modeling”, Transactions on Information Theory, pp. 25–28, 1987.

[21] Liu, C. S., Lee, C. H., Chou, W., Juang, B. H. and Rosenberg, A. E., “A study on minimum error discriminative training for speaker recognition”, The Journal of the Acoustical Society of America, pp. 637–648, 1995.

[22] Dimopoulos, S., Potamianos, A., Lussier, E., and Lee, C., “Multiple Time Resolution Analysis of Speech Signal using MCE Training with Application to Speech Recognition”, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3801-3804, 2009.

[23] Juang, B. H., Hou, W. and Lee, C. H., “Minimum classification error rate methods for speech recognition”, IEEE Transactions on Speech and Audio Processing, pp. 257–265, 1997.

[24] Schmitt, A., Ultes, S. and Minker, W, “A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let’s Go Bus Information System”, LREC, pp. 3369–3373, 2012.

[25] SpeDial Project, “SpeDial Project free data deliverable D2.1.”, https://sites.google.com/site/spedialproject/risks-1, 2013–2015.

[26] Z.-H. Tan and B. Lindberg, “Low-complexity variable frame rate analysis for speech recognition and voice activity detection.” IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pp. 798-807, 2010.

[27] Hall, M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., “The WEKA Data Mining Software: An Update”, SIGKDD Explorations, vol. 11, 2009