Latent Semantics as Cognitive Components

Michael Kai Petersen, Morten Mørup and Lars Kai Hansen

DTU Informatics, Cognitive Systems, Technical University of Denmark
Building 321, DK-2800 Kgs.Lyngby, Denmark

{mkp,mm,lkh}@imm.dtu.dk

Abstract— **Cognitive component analysis, defined as an unsupervised learning of features resembling human comprehension, suggests that the sensory structures we perceive might often be modeled by reducing dimensionality and treating objects in space and time as linear mixtures incorporating sparsity and independence. In music as well as language the patterns we come across become part of our mental workspace when the bottom-up sensory input raises above the background noise of core affect, and top-down trigger distinct feelings reflecting a shift of our attention. And as both low-level semantics and our emotional responses can be encoded in words, we propose a simplified cognitive approach to model how we perceive media. Representing song lyrics in a vector space of reduced dimensionality using LSA, we combine bottom-up defined term distances with affective adjectives, that top-down constrain the latent semantics according to the psychological dimensions of valence and arousal. Subsequently we apply a Tucker tensor decomposition combined with re-weighted l_1 regularization and a Bayesian ARD automatic relevance determination approach to derive a sparse representation of complementary affective mixtures, which we suggest function as cognitive components for perceiving the underlying structure in lyrics.**

I. Introduction

Trying to make sense of the world, whether combining segments of lines and edges into visual scenes or assembling soundscapes from rhythmical and spectral contours of pitch, we perceptually search for patterns in sensory data that spatially seem to form clusters or sequentially re-occur in time. Either based on how frequently something happens infer the likelihood of it occurring within a certain structure. Or by grouping sensory inputs into larger gestalts, that we perceive as a whole to organize them in the simplest possible way. Likelihood can be seen as just another way of defining simplicity [1], and we would thus also expect cognitive models to be optimized for providing the briefest description of features underlying the probabilistic hierarchical structures we perceive [2]. While the features of images or sounds may constitute complex manifolds in a high-dimensional space, the sensory coding in the brain takes advantage of the underlying regularities and transforms sensation into sparse representations [3]. But encoding a state space with more vectors than we have input dimensions available requires that we have additional information on what is the underlying structure. Allowing us to reconstruct the original signal from a incomplete set of linear mixtures, using an optimization approach equivalent to l_0 norm regularization, where insignificant components are reduced to zero and only the essential features in the input are retained [4]. In essence flattening the manifolds that we perceive by hierarchically aligning our receptive fields to the structure in the input [5]. Or in other words, semantics emerge when exploiting a maximally compressed description of what we perceive [6]. Language itself functions as such a piece of self-replicating code, which based on hierarchically nested constructs and spatiotemporal constraints generate patterns allowing us to recursively encode new concepts [7].

It has earlier been shown that COCA cognitive component analysis, defined as an unsupervised learning of features resembling how we perceive sensory structures, might enable machine learning classification of musical genres [8] or phonemes in speech processing [9], based on ICA independent component analysis [10]. Or similarly using a sparse constrained NMF non-negative matrix factorization, allow for retrieving a part-based representation of facial features from a linear mixture of statistically independent contexts [11]. However, such a retrieval of independent components may only partly resemble how we cognitively overcome challenges like the `cocktail party problem' of carrying on a conversation threatened to be overshadowed by other voices [12], as our brains are able to pick out a stream of particular interest based on embodied cognitive processes boosting the signal to noise ratio. In essence by top-down applying selective attention to switch between cross-correlated segregated features, which when subsequently grouped form a perceived outline similar to what makes a figure stand out from the background in an image [13].

Both in music and language the patterns we perceive become part of our mental workspace when the bottom-up sensory input raises above the background noise of core affect [14], and top-down trigger distinct feelings reflecting a shift of our attention [15]. And as both low-level semantics and our emotional responses can be encoded in words, we propose a simplified cognitive approach to model how we perceive media based on texts associated with the content. Here exemplified by a large selection of song lyrics that are represented in a vector space of reduced dimensionality. Using LSA latent semantic analysis [16], we bottom-up define term distances between the words in the lyrics and a selection of affective adjectives, that top-down constrain the latent semantics according to the psychological dimensions of valence and arousal. Subsequently we apply a multi-way Tucker tensor decomposition to the LSA matrices and as a result derive a part-based sparse representation of complementary affective mixtures and temporal components, which we propose might interact as cognitive components for perceiving the emotional structure in media.

II. RELATED WORK

Advances in neuroimaging technologies that enable studies of brain activity have established that musical structure to a larger extent than previously thought is being processed in `language' areas of the brain [17]. Specifically related to songs, fMRI `functional magnetic resonance imaging' experiments show that neural processes involved in perception and action when covertly humming the melody or rehearsing the song text activate overlapping areas in the brain. This indicates that core elements of lyrical music appear to be treated in a fashion similar to those of language [18], which is in turn supported by EEG `electroencephalograhy' studies showing that language and music compete for the same neural resources when processing syntax and semantics [19]. Looking specifically into the functional architecture of memory, it appears that both storage and representation of verbal and tonal information rely on the same neural networks. That is, processing and encoding of phonemes as well as pitch are largely based on the same sensorimotor mechanisms. The phonological loop which stores words for a few seconds in working memory when we subvocally repeat syllables [20], appear not only to be used in speech but similarly involved when maintaining a sequence of tones in memory [21]. Studies of the interaction between phonology and melody indicate that ``vowels sing whereas consonants speak", meaning that vowels and melodic intervals may have similar functionalities related to the generative structure of syntax in language and music involving both hemispheres of the brain [22]. While experiments investigating whether tunes are priming texts or the other way around, suggest that lyrics and melody are mutually accessible in song memory based on a symmetrical two-way relationship [23].

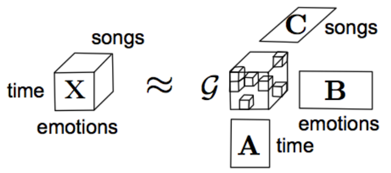
Cognitively speaking our feelings can be thought of as labels that are consciously assigned to the ebb and flow of emotions triggered by sensory inputs [24]. That is, the brain applies an 'analysis-by-synthesis' approach, which infers structure from bottom-up processing of statistical regularities, that are continuously compared against stored patterns of top-down labeled gestalts [25]. Language builds on sensory-motor mechanisms in the brain, which fuses the various modalities of sound, sight or sensations of touch and texture together in a semantic structure of action concepts. Reading a word like `smile' triggers the same motor resonances in our brains as a visual representation of the corresponding facial features, which means that also verbal emotional expressions are embodied [26]. Experiments exploring how we perceive emotions have shown that while we often think of affective terms as describing widely different states, these can be represented as related components in a circumplex model framed by the two psychological primitives: valence and arousal [27]. Within this emotional plane the dimension of valence describes how pleasant something is along an axis going from positive to negative contrasting words like 'happy' against 'sad', whereas arousal captures the amount of intensity ranging from passive states like 'sad' to aspects of excitation reflected in terms like 'angry' or 'funny'. This mapping of feelings has actual neural correlates, as brain imaging studies using fMRI to trace which parts become involved when people read emotional words, indicate that activation is divided into two distinct neural networks which are linearly correlated with the values of valence or arousal [28]. Even though `happy' and `sad' are placed at the far ends of the two axes of valence and arousal, these feelings are often perceived at the same time in psychological experiments and should not be considered bipolar opposites, but rather as interchanging in rapid succession of each other. Similar to how we perceive the parallel lines formed by a three dimensional Necker cube like two different objects, as we alternate between seeing the same shape in a perspective viewed from either the top or the bottom [29].

III. METHOD

In our proposed cognitive model the bottom-up generated sensory data is a matrix of rows and columns representing the occurrences of words within multiple contexts. The foundation here is a large text corpus which allows for modeling the terms as linear combinations of the multiple paragraphs and sentences they occur in. The underlying text corpus is based on 22829 terms found in 67380 contexts, divided into 500 word segments, made from 22072 literature and poetry excerpts of the *Harvard Classics*, 15340 segments of *Wikipedia* music articles, and 29968 general news items from the *Reuters Corpus* gathered over the period 1996-1997. Our analysis is based on 50.274 lyrics selected from *LyricWiki* by using artist entries retrieved from the *Wikipedia* ``List of Musicians", associated with the genres: alternative rock, blues, brit pop, dream pop, gothic rock, indie rock, indie pop, pop punk, R&B, soul, hard rock, reggae and heavy metal. When we project the lyrics into the LSA space the aim is to find matrices of lower dimensionality embedding the underlying structures, which when multiplied allow us to reconstruct the original sensory data. These higher order associations within the original matrix emerge as similar features appearing in a large number of contexts that are simultaneously squeezed into a reduced number of rows and columns that correspond to orthogonal directions capturing the highest variance in the data based on SVD singular value decomposition [30]. Capturing what constitutes the highest variance in a new set of variables similar to PCA principal components, earlier studies of emotional words have shown that the first component alone would reflect almost half of the total variance based on contrasts between 'happy' and 'sad'. Whereas juxtapositions of frustration against tranquility, and aspects of negative arousal can be accounted for by the second and third PCA component respectively [27].

Projecting the lyrics into the LSA semantic space we define the cosine similarity between vectors representing the individual lines making up each of the lyrics against twelve affective adjectives: 'happy, funny, sexy, romantic, soft, mellow, cool, angry, aggressive, dark, melancholy, sad'. The twelve affective adjectives, that in our model emulate how top-down aspects of attention trigger distinct feelings in response to evoked emotions, have been selected among the terms most frequently applied as emotional tags by users describing music in the *last.fm* social network [31]. And additionally represent contrasting aspects of valence and arousal that for most of the adjectives are defined based on user rated values along the two dimensions assessing how the terms are being perceived [32]. To determine the optimal number of factors when reducing the original term document matrix we submit the LSA setup to a TOEFL ‘test of english as a foreign language’ while varying the number of dimensions until an optimal percentage of correct answers are returned. For our LSA setup the best fit corresponds to 71,25 % correctly identified synonyms in the TOEFL test when reducing the matrix to 125 factors, thus providing a result above the 64.5 % average achieved by non-native college applicants, as well as the 64.5 % and 70.5 % correct answers previously reported for LSA spaces and probabilistic LDA topic models respectively [33]. To provide an additional measure of ground truth we have earlier extracted LSA emotional topics over time from 24798 lyrics, and compared the resulting patterns (Fig,1) against user-defined tags describing the corresponding songs at *last.fm* [34]. Although the *last.fm* tags describe an entire song whereas the LSA provides a time-series analysis of the lyrics line by line, correlations were found between `happy-sad' emotions, as well as aspects defining `soft, cool' and `dark' textures. Although it is feasible to model audio features as a spectral `bag of frames' for urban soundscapes like kids playing in a park, this is not the case for somebody playing a solo violin partita by Bach. In music a minority of frames are statistically insignificant outliers providing the underlying semantic structure [35]. Features forming musical phrases are not distributed as random segments, but form patterns related to perceptually significant peaks or local maxima that generate the larger scale semantic structures [36].

In order to explore what affective and temporal components might cognitively enable us to encode the manifolds of lyrics as sparse representations of features aligned to the structure in the input, we move in our analysis beyond the initial LSA second order matrices. Subsequently we apply a three-way Tucker tensor model [37] in order to find what factors are significantly correlated within the LSA matrices when compared across the fifty thousand lyrics selected from *LyricWiki*. To enable a comparison of the songs independent of duration the LSA matrices were resampled to a fixed length of 32 time points, corresponding to the average number of lines in the lyrics. Decomposing the LSA derived emotions over time patterns into a three dimensional tensor, makes it possible to assess the strengths by which the vector loadings of the time and emotions matrices interact over a large number of songs:



$$x_{ijk} \approx (\mathcal{G} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C})_{ijk} = \sum_{lmn} g_{lmn} a_{il} b_{jm} c_{kn}$$

where the core array $\mathcal{G}$ is positive and defines the strength by which the columns or vector loadings of the **A** *Time x L* (positive), **B** *Emotions x M* (unconstrained), and **C** *Song x N* (positive) matrices interact. Meaning, the model captures all potential linear interactions between emotions, time and songs. And the variables *L, M* and *N* correspond to the number of components or columns in the factor matrices **A**, **B** and **C**, which could in turn be interpreted as principal components in each of the three modes [38]. To assure that the model provides the most sparse representation, the Tucker tensor decomposition is fitted using a sparse regression algorithm, where excess components are pruned by regularization based on the l_1 norm to minimize non-zero elements in the core array. What components would be necessary for representing the interactions between the three modalities, or conversely which ones could be set to zero, depend on the amount of regularization defined using a hierarchical Bayesian ARD automatic relevance determination approach. Learning the hyperparameters of the priors based on the relevancy of features in the data, the Bayesian ARD defines a range of variation for the underlying parameters. Providing a sparse representation, these parameters are modeled as the width of exponential and Laplace prior distributions, which are non-negativity constrained and unconstrained respectively assigned to the loadings and core. To reduce the interaction between components expressed in the core we evaluate the relevance by Bayesian ARD of each core element separately rather than imposing an equal degree on all core elements as proposed in [39]. As a result, the corresponding model is specified by:

$$\begin{aligned}
P(\boldsymbol{a}_l|\lambda_l^a) &= (\lambda_l^a)^I e^{-\lambda_l^a |\boldsymbol{a}_l|_1}, \quad s.t. \quad a_{il} > 0 \\
P(\boldsymbol{b}_m|\lambda_m^b) &= \frac{(\lambda_m^b)^J}{2} e^{-\lambda_m^b |\boldsymbol{b}_m|_1} \\
P(\boldsymbol{c}_n|\lambda_n^c) &= (\lambda_n^c)^K e^{-\lambda_n^c |\boldsymbol{c}_n|_1}, \quad s.t. \quad c_{kn} > 0 \\
P(g_{lmn}|\lambda_{lmn}^g) &= \lambda_{lmn}^g e^{-\lambda_{lmn}^g |g_{lmn}|_1}, \quad s.t. \quad g_{lmn} > 0 \\
P(\lambda_t^q|1,\epsilon) &\sim Gam(\lambda_t^q|1,\epsilon)
\end{aligned}$$

Taking the logarithm we estimate all model parameters by maximum likelihood of the posterior log likelihood function $\log L$:

$$\begin{aligned}
\log L \quad \propto \quad & \text{const.} - \frac{1}{2\sigma^2} \|\mathcal{X} - \mathcal{G} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}\|_F^2 \\
- \quad & \sum_l \lambda_l^a (|\boldsymbol{a}_l|_1 + \epsilon) + I \sum_l \log \lambda_l^a \\
- \quad & \sum_m \lambda_m^b (|\boldsymbol{b}_m|_1 + \epsilon) + J \sum_m \log \lambda_m^b \\
- \quad & \sum_n \lambda_n^c (|\boldsymbol{c}_n|_1 + \epsilon) + K \sum_n \log \lambda_n^c \\
- \quad & \lambda_{lmn}^g (|g_{lmn}|_1 + \epsilon) + \sum_{lmn} \log \lambda_{lmn}^g
\end{aligned}$$

As such, each alternating optimization problem of the model parameters $\mathcal{G}$, **A, B, C** form a standard l_1 regularized regression problem:

$$\frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{Q}\boldsymbol{s}\|_F + \sum_j \lambda_j |s_j|_1$$

whereas the ARD hyperparameters are updated according to:

$$\begin{aligned}
\lambda_l^a &= \tfrac{I}{|\boldsymbol{a}_l|_1 + \epsilon}, & \lambda_m^b &= \tfrac{J}{|\boldsymbol{b}_m|_1 + \epsilon}, \\
\lambda_n^c &= \tfrac{K}{|\boldsymbol{c}_n|_1 + \epsilon}, & \lambda_{lmn}^g &= \tfrac{1}{|g_{lmn}|_1 + \epsilon}.
\end{aligned}$$

We adjusted ε and determined σ^2 such that the signal to noise ratio is 0 dB, thus not assuming more signal than noise when applying the sparse Bayesian algorithm [39]. Each of the hyperparameters are updated according to the norm of their respective component, and as such the updates of the core elements are equivalent to l_0 optimization based on re-weighted l_1 norm optimization [40]. To optimize the likelihood function, the combined Tucker ARD approach was applied ten times to the LSA matrices, and the decomposition achieving the highest logarithmic probability value based on 1000 iterations was selected to provide the best representation of the data.

IV. Results

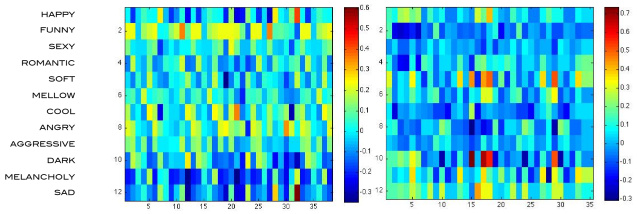


Fig. 1. LSA lyrics matrices of feelings triggered over time, exemplified by Amy Winehouse's funny and cool "Rehab" (left) whereas Evanescence's "My Immortal" (right) bring out soft and dark. The overall emotions of the corresponding songs are tagged "mellow, sexy, cool, happy" and "mellow, sad, melancholy, romantic" by users in the *last.fm* social network.

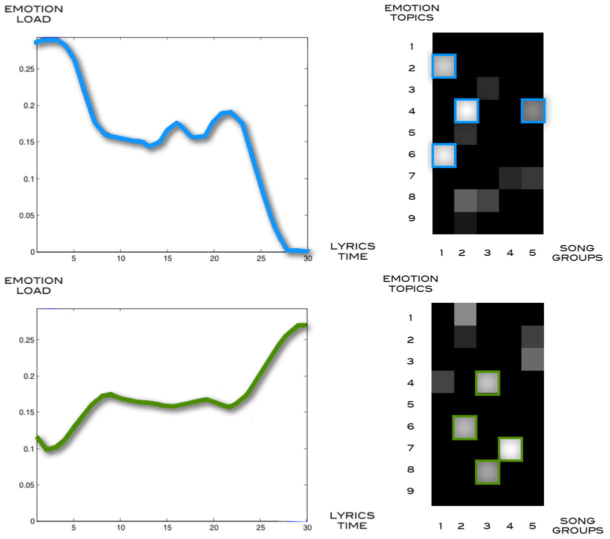


Fig. 2. Correlated components among emotions, time and songs in the tensor core array, identified in a 3-way sparse ARD Tucker decomposition of LSA matrices representing 50.274 lyrics. The saturated cells outlined in blue and green (top and bottom) within the core array, indicate that the maximum interaction is concentrated in five emotional topics (2, 4, 6, 7 & 8) within five groups of songs. The loadings of these components in the core array are in turn related to the two time series curvatures representing the variability of emotional load over time outlined in blue and green (top and bottom).

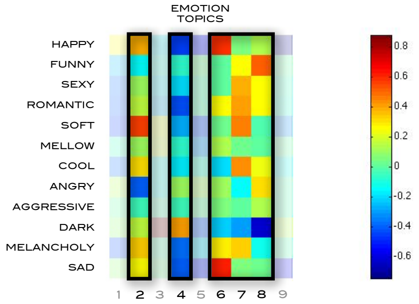


Fig. 3. Emotional topics constituted by mixtures of 2.'soft', 4. 'dark' 6.'happy-sad', 7, 'soft-cool' and 8.'funny-angry' feelings.

V. Discussion

When projecting song texts into the LSA space by defining term vector distances in relation to the selected emotional adjectives, the columns of the matrices reflect a vertical span from positive to negative valence. Taking the song "Rehab" (Fig.1) as an example, the upper half of the lyrics matrix is in row 2 characterized by a sustained band of `funny' coupled with activations of 'cool' and 'angry' components in row 7 and 8. The rows of 'happy' and 'sad' emotions at the very top and bottom remain inactive until being triggered towards the very end. Whereas in "My immortal" (Fig.1) the lyrics trigger the bottom rows 10-12 of the matrix, reflecting mostly 'dark' as well as 'melancholy' and 'sad' components, while the upper part of the matrix remains inactive. Such LSA patterns may capture the overall emotional bias towards 'happy' or 'sad' reflected in the user defined tags describing the corresponding songs at *last.fm.* But also suggests that the feelings continuously change over time, and might be generated from a few significant peaks rising above the affective building blocks forming the contrasting emotional patterns.

Applying a three-way ARD Tucker tensor decomposition to the LSA matrices based on a large sample of lyrics, the sparse core array indicates that two time-series components capture what constitutes the emotional load over the duration of a song. The first time-series component outlined in blue forms a descending curve (Fig.2), correlated with the emotional topics 2, 4 and 6 representing mixtures of 'soft' and 'dark' textures as well as 'happy-sad' contrasts (Fig.3). While the second time-series component outlined in green is an ascending line (Fig.2), associated with the emotional topics 4, 6 7 and 8, which besides 'dark' and 'happy-sad' mixtures also represent 'soft-cool' textures and a more aroused 'funny-angry' topic (Fig.3). And in the Tucker core array these time-series and emotional components come out as correlated with five groups of songs.

Looking into top samples from song group 1, exemplified by Bon Jovi's "Not fade away" and The Mission's "Love", the saturated juxtaposition of 'happy-sad' found in emotion topic 6 can be made out in both lyrics (Fig.4). Taking two of the top tracks representative of song group 2 as examples, Nirvana's "Love Buzz" and the Therapy? song "Stay Happy" again reflect the simultaneous coupling of 'happy-sad' but less sustained and in the latter song biased towards 'happy' (Fig. 5).

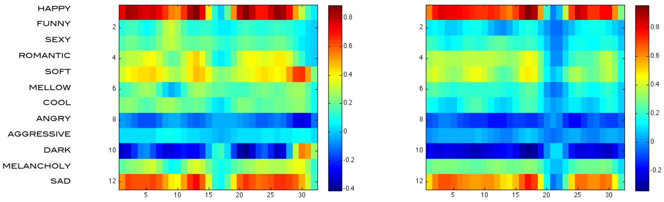


Fig. 4. Song group 1 correlated with emotional topics 2 and 6, exemplified by the lyrics of Bon Jovi's "Not fade away" (left) and The Mission's rendering of "Love" (right) .

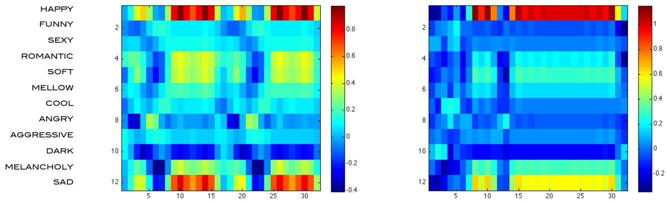


Fig. 5. Song group 2 correlated with emotion topics 4 and 6, exemplified by the lyrics of Nirvana's "Love Buzz" (left) and the Therapy? song "Stay Happy" (right) .

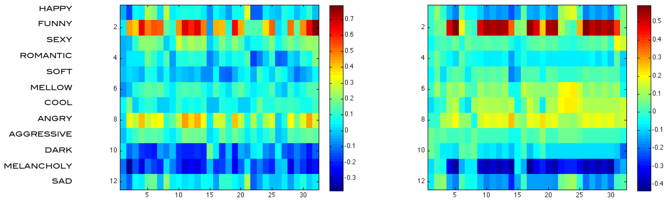


Fig. 6. Song group 3 correlated with emotion topics 4 and 8, exemplified by the lyrics in The Sex Pistol's "No Fun" (left) and the Michael Jackson song "Jam" (right)

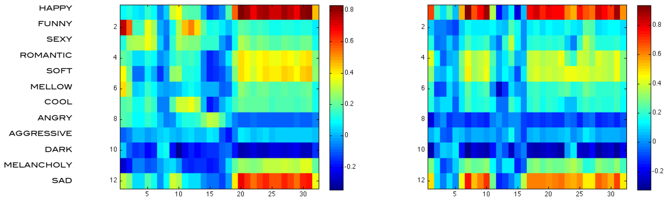


Fig. 7. Song group 4 correlated with emotion topic 7, exemplified by the lyrics of Bo Didley's "Diddley Daddy" (left) and the Lou Reed song "The Blue Mask Women" (right)

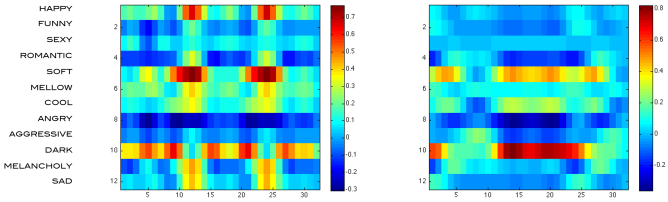


Fig. 8. Song group 5 correlated with emotion topic 4, exemplified by the lyrics of The Doors's "End of the night" (left) and the Yeah Yeah Yeah's song "Hello Tomorrow" (right).

In song group 3, exemplified by some more problematic lyrics: the The Sex Pistols' "No Fun" and the Michael Jackson song "Jam" (Fig.6), strongly reflect emotion topic 8 capturing 'funny-angry' aspects in the lyrics. However both highlight the problem of treating lyrics as a bag of words, as the former song laments being alone, while the latter seems rather to channel energetic aspects of arousal than being 'funny' as such. Song group 4 is mainly representing the characteristics of emotion topic 7, which captures 'romantic-soft' aspects as in the plots of Bo Didley's "Diddley Daddy" and the Lou Reed song "The Blue Mask Women" (Fig.7). Although again the lyrics simultaneously trigger 'happy-sad' contrasts, making emotion topic 7 seem more like a complementary texture than a principal emotional component. Also the top samples of song group 5 activate 'soft-dark' textures that can metaphorically be interpreted as feelings (Fig.8). Together with the previously identified 'happy-sad' and 'funny-angry' mixtures, these components appear to define the emotional building blocks of the lyrics.

The sparse representation of emotional topics can largely be interpreted as contrasts spanning the psychological axes of valence and arousal that in the widest sense define a low dimensional representational structure of the input. In turn reflecting recent neuroscientific findings indicating that the processing of emotional words appear literally divided among two distinct neural networks, linearly correlated with the values of valence or arousal [28]. Interpreted as principal components the 'happy-sad' mixture could be thought of as representing the maximal contrast potentially biased towards positive or negative valence, which as complementary aspects define the emotional range. Whereas aspects of excitation reflected in the emotional mixture of `funny-angry' seem to capture the amount of arousal as perceived intensity. So, together these two pairs of contrasts might be interpreted as representative of the two principal dimensions framing a psychological space that provides the constraints for how we encode emotions. Also the identified `soft', `cool' or `dark' textures identified in the lyrics appear salient as they might not only be understood as abstract concepts, but in a larger context reflect somatosensori aspects of touch or timbre which are metaphorically mapped onto feelings as has previously been documented [15]. In essence the sparse representation of emotional topics and time-series curvatures of emotional load retrieved in our analysis are derived from almost 20 million free variables originally defining the 12 dimensional vectors of affective adjectives within 50274 matrices. When applying the combined ARD Tucker tensor decomposition to the LSA matrices the problem space is reduced by a factor of 77 to 251.632 variables, while the degree of explanation retained in the model remains 21 %. Whereas a null hypothesis for the Tucker model based on a random permutation of the data would account for only 4.84 +/- 0.01 % of the variance.

VI. Conclusion

Cognitive component analysis, defined as an unsupervised learning of features resembling human comprehension, suggests that the sensory structures we perceive might often be modeled by reducing dimensionality and treating objects in space and time as linear mixtures incorporating sparsity and independence. However, such compressed representations may only partly resemble how we cognitively perceive media, if combining the bottom-up inferred patterns of features co-occurring in multiple contexts, with top-down aspects of attention reflecting our conscious emotional responses to what we encounter. As both low-level semantics and our emotional responses can be encoded in words, we propose a simplified cognitive approach to model how we perceive media based on texts associated with the content. Deriving a part-based sparse representation of complementary affective mixtures and temporal curvatures, we propose that these might interact as cognitive components enabling us to perceive the emotional structure in media. Constrained to the two psychological dimensions of valence and arousal, the combinations of emotional topics could provide the contrasting elements that define a low dimensional representational structure of the input. Which would allow us to reconstruct the original signal from a incomplete set of linear affective mixtures forming sequential patterns that temporally reflect the emotional load. Embedded as cognitive components that we are able to retrieve as latent semantics when the bottom-up generated input raises above the background noise of core affect and top-down trigger distinct feelings in response to what we perceive.

REFERENCES

[1] P. Vitanyi and M. Li, *Minimum description length induction, Bayesianism, and Kolmogorov complexity*, IEEE Transactions on Information Theory, 46:2, 2000.

[2] N. Chater and G. Brown, *From universal laws of cognition to specific cognitive models*, Cognitive Science, 32, 2008.

[3] B.A. Olshausen and D.J.Field, *Sparse coding of sensory inputs*, Current Opinion in Neurobiology, 14, 2004

[4] E.J. Candes, M.B. Wakin and S.P. Boyd, *Enhancing sparsity by reweighted l_1 minimization*, Journal of Fourier Analysis and Applications, 14, 2008.

[5] H.B. Barlow, *Single units and sensation: a neuron doctrine for perceptual psychology*? Perception, 1, 1972.

[6] E. Baum, *What is thought ?*, MIT Press, 2004.

[7] G. Lakoff, and M. Johnson, *Philosophy in the flesh: the embodied mind and its challenge to western thought,* Basic Books, 1999

[8] L.K. Hansen, P. Ahrendt and J. Larsen, *Towards cognitive component analysis*, Proceedings of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, 2004.

[9] L. Feng and L.K. Hansen, *On phonemes as cognitive components of speech*, IAPR workshop on cognitive information processing, 2008.

[10] P. Comon, *Independent component analysis, a new concept?*, Signal Processing, 36, 1994.

[11] M. Mørup, K.H. Madsen and L.K. Hansen, *Approximate Lo constrained non-negative matrix and tensor factorization,* Proceedings of Circuits and Systems, ISCAS, DOI: 10.1109/ISCAS.2008.4541671, 2008.

[12] S. Haykin and Z. Ghen, *The cocktail party problem,* Neural Computation, 17, 2005.

[13] Z. Ghen, *Stochastic correlative firing for figure-ground segregation,* Biological Cybernetics, 92, 2005.

[14] S. Dehaene and L. Naccache, *Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework,* Cognition, 79, 2001

[15] A. Damasio, *Feelings of emotion and the self,* Annals of the New York Academy of Sciences, 1001, 2003

[16] T.K. Landauer and S.T. Dumais, *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge,* Psychological Review, 104:2, 1997

[17] A.D. Patel, *Music, language and the brain,* Oxford University Press, 2008

[18] D.E. Callan et al., *Song and speech: Brain regions involved with perception and covert production,* NeuroImage, 31:3, 2006

[19] S. Koelsch, *Neural substrates of processing syntax and semantics in music,* Current Opinion in Neurobiology, 15, 2005

[20] A. Baddeley, *Working memory: looking back and looking forward,* Nature Neuroscience, 4, 2003

[21] S. Koelsch, *Functional architecture of verbal and tonal working memory: an fMRI study,* Human Brain Mapping, 30:3, 2009

[22] R. Kolinsky et al., *Processing interactions between phonology and melody: Vowels sing but consonants speak,* Cognition, 112, 2009

[23] I. Peretz, M. Radeau and M. Arguin, *Two-way interactions between music and language: Evidence from priming recognition of tune and lyrics in familiar songs,* Memory & Cognition, 32:1, 2004

[24] A. Damasio, *The feeling of what happens: body, emotion and the making of consciousness,* Vintage, 1999

[25] E. Borenstein and S. Ullman, *Combined top-down / bottom-up segmentation,* IEEE Transactions on pattern analysis and machine intelligence, 30:12, 2008

[26] F. Foroni and G. Semin, *Language that puts you in touch with your bodily feelings,* Psychological Science, 20:8, 2009

[27] J.A. Russell, *A circumplex model of affect,* Journal of Personality and Social Psychology, 39:6, 1980

[28] J. Posner et al., *The neurophysiological bases of emotion: an fMRI study of the affective circumplex using emotion-denoting words,* Human Brain Mapping, 30, 2009

[29] L.F. Barrett and E. Bliss-Moreau, *Affect as psychological primitive,* Advances in experimental social psychology, Burlington Academic Press, 2009

[30] G.W. Furnas et al., *Information retrieval using a singular value decomposition model of latent semantic structure,* Proceedings of 11th annual international ACM SIGIR conference, 1988

[31] X. Hu, M. Bay and S. Downie, *Creating a simplified music mood classification ground-truth set,* Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR, 2007

[32] M.M. Bradley and P. Lang, *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings,* The Center for Research in Psychophysiology, University of Florida, 1999

[33] T.L. Griffiths, M. Steyvers and J.B. Tenenbaum, *Topics in semantic representation,* Psychological Review, 114:2, 2007

[34] M.K. Petersen and L.K. Hansen, *Modeling lyrics as emotional semantics,* Proceedings of YoungCT, KAIST Korea Advanced Institute of Science and Technology, 2010

[35] M. Aucouturier, B. Defreville and F. Pachet, *The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music,* Acoustical Society of America, 122:2, 2007

[36] M. Levy and M. Sandler, *Music information retrieval using social tags and audio,* IEEE Transactions on Multimedia, 11:3, 2009

[37] L.R. Tucker, *Some mathematical notes on three-mode factor analysis,* Psychometrika, 31:3, 1966

[38] T.G. Kolda and B.W. Bader, *Tensor decompositions and applications,* SIAM Review, June, 2008

[39] M. Mørup and L.K. Hansen, *Automatic relevance determination for multi-way models,* Journal of Chemometrics, 51:3, 2009

[40] E.J. Candes, M.B. Wakin and S.P. Boyd, *Enhancing sparsity by reweighted L1 minimization,* Journal of Fourier Analysis and Applications, 23:7-8, 2008