TASK-DRIVEN ATTENTIONAL MECHANISMS FOR AUDITORY SCENE RECOGNITION

Kailash Patil and Mounya Elhilali

Center for Language and Speech Processing, Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD, USA.

kailash@jhu.edu, mounya@jhu.edu

ABSTRACT

How do humans attend to and pick out relevant auditory objects amongst all other sounds in the environment? Based on neurophysiological findings we propose two task oriented attentional mechanisms acting as Bayesian priors which act on two separate levels of processing: a sensory mapping stage and object representation stage. The former sensory stage is modeled as a high dimensional mapping which captures the spectrotemporal nuances and cues of auditory objects. The latter object representation stage then captures the statistical distribution of the different classes of acoustic scenes. This scheme shows a relative improvement in performance by 81% compared to a baseline system.

Index Terms— Auditory Attention, Acoustic Scene Analysis, Sensory Processing, Object based attention.

1. INTRODUCTION

An auditory object is often equated to the sound produced by a single source [1]. While the correspondence between the two is not always a one-to-one mapping, the soundscape incident on a listener generally consists of multiple auditory objects that constitute the acoustic scene. Identifying an auditory object is not a trivial task, especially since each object can present itself in a multitude of variations. For example, the blast of a car horn can differ depending on the make, the speed of the vehicle, and also the distance of the vehicle from the listener. Subsequently, this makes the identification of a collection of auditory objects (i.e. the acoustic scene) significantly harder. To add to this complexity, the nature and number of acoustic objects in a typical scene change over time. In a street, the sounds coming from passing cars can blend every now and then with speech from pedestrians or music from street artists. Changing scenarios add a new dimension of difficulty to the task of acoustic scene classification.

Attempts at automatic acoustic event and scene classification have typically followed the path of extracting short term features from waveforms and learning the statistics of these features to later classify an unknown example. Mel Frequency Cepstral Coefficients (MFCC), filterbank energies or Perceptual Linear Prediction coefficients (PLP) have been popularly used as features for this task [2, 3, 4]. They are often complimented with other low level features like zero crossing rate, short time energy, spectral flux, pitch, brightness and bandwidth [3, 5, 6] or are transformed to account for long term statistics [7]. Though short term spectral attributes coupled with low-level features have been quite successful in a number of applications, studies have also shown that they are limited in capturing the full range of information relevant for acoustic scene recognition; and that joint local modulations in energy along both time and frequency are able to better capture the qualities of acoustic scenes [8]. This rich modulation space builds on neurophysiological studies in the mammalian auditory system indicating that neurons at the level of auditory cortex respond to local joint spectral and temporal modulation in the signal [9]. This biological analysis can be viewed as mapping sound onto a high dimensional feature space which captures the detailed variations of the spectral profile and its temporal variations, as a basis for representing acoustic events.

This sensory mapping is complemented with cognitive mechanisms, most notably task-driven attention, which allows us to isolate and recognize objects of interest amidst other competing sound events [10]. Neurophysiological and brain imaging studies have shown that task-driven attention modulates the gain of sensory cortex responses to highlight features of interest [11, 12]. Attention has been argued to act as a Bayesian prior representing distribution of beliefs acting as gating mechanism to reduce uncertainty, to increase signal-to-noise ratio or to refine perceptual inference around some goal-specific point in sensory space [13]. In addition, attention is also believed to modulate cognitive and decision-making frontal areas of the brain, most notably prefrontal cortex [14]. Psychoacoustic evidence also supports the premise that attention operates at multiple levels, be it feature-based or object-based levels of representation [15, 16, 17, 18]. Motivated by these observations, the current study attempts to develop a scheme that incorporates attentional mechanisms in a model for scene recognition for multi-source environments. The model focuses on attentional processes operating at the level of both sensory representation and cognitive decisions.

This work was partly supported by grants from NSF (IIS-0846112), NIH (1R01AG036424) and ONR (N000141010278 and N00014-12-1-0740).

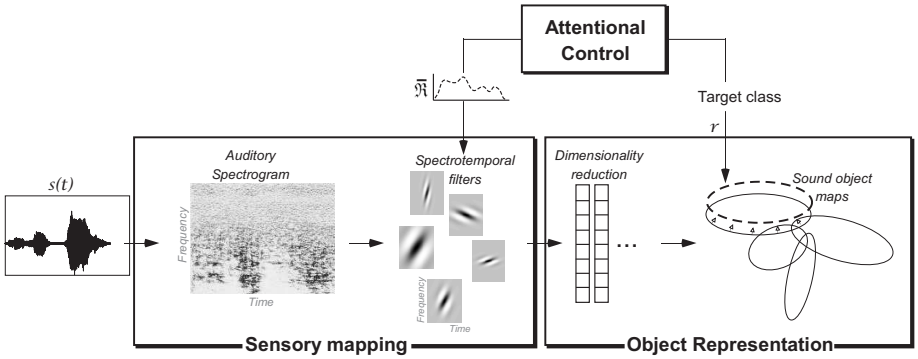

Fig. 1. Schematic of the proposed model used for the task of scene classification. Attention could be applied both at the sensory stage and the object recognition stage.

2. METHODS

The proposed model is divided into Sensory Processing, Object Representation and Adaptation modules as shown in Fig1. Each of these modules and the experimental setup is described below.

2.1. Sensory Processing

The incoming sound is processed to extract informative features using techniques that mimic the behavior of the mammalian auditory system. This can be further divided into two steps - the subcortical stage and the cortical processing stage. In the subcortical stage, the waveform is passed through a set of 128 asymmetric filters $h(t;f)$ placed uniformly on a logarithmic axis covering 5.3 octaves starting from $180Hz$. This is similar to the frequency-space transformation of the cochlear membrane. This is followed by a spectral derivative and a half wave rectification stage, which models the lateral inhibition networks in the cochlear nucleus, sharpening the frequency resolution of these filters. The mid brain processing is implemented as a short term integration with window $\mu(t;\tau) = e^{-t/\tau}u(t)$ and $\tau = 2ms$ followed by cubic root compression. These subcortical transformations can be collectively written as in Eq. 1 and the details of implementation can be found in [19].

$$y(t,f) = (max(\partial_f(s(t) \otimes_t h(t;f)), 0) \otimes_t \mu(t;\tau))^{\frac{1}{3}} \quad (1)$$

where $\otimes_t$ represents convolution with respect to time.

This resulting time-frequency representation is referred to as the auditory spectrogram. In the cortical stage, this spectrogram is analyzed locally for joint spectrotemporal modulations using a bank of modulation tuned filters. These filters as defined in Eq. 2, are shaped like 2D Gabors, which are known to be a linear approximation to the receptive field shapes of auditory cortex neurons [20, 21]. The temporal modulation rate and spectral modulation rate are denoted by $\mathfrak{r}$ and $\mathfrak{s}$ respectively. The filtering operation can then be written as simple two dimensional convolution as in Eq. 3 which yields a four dimensional tensor representation.

$$MF(f,t;\mathfrak{s},\mathfrak{r}) = \frac{1}{2\pi\sigma_t\sigma_f} e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_t^2}+\frac{f^2}{\sigma_f^2}\right)} e^{2\pi i(\mathfrak{r}t+\mathfrak{s}f)} \quad (2)$$

$$\mathfrak{R}(f,t;\mathfrak{s},\mathfrak{r}) = |y(f,t) \otimes_{f,t} MF(f,t;\mathfrak{s},\mathfrak{r})| \quad (3)$$

The MF filters are tuned to 10 upward rates and 10 downward rates $\{\mathfrak{r}$ = 2, 3.4, 5.7, 9.5, 16, 26.9, 45.3, 76.1, 128, 215.3 Hz$\}$ and 11 scales $\{\mathfrak{s}$ = 0.25, 0.35, 0.5, 0.71, 1, 1.41, 2, 2.83, 4, 5.66, 8 cycles/octave$\}$, resulting in a total of 220 filters.

2.2. Object Representation

Each audio recording is windowed into non-overlapping 1s segments. We integrate the cortical representation $\mathfrak{R}$ over the time duration of each window. To facilitate the machine learning module we reduce the number of dimensions via Tensor Singular Value Decomposition [22] to keep 99% of the variance resulting in a 336 dimensional feature vector. We learn the distribution of these feature vectors for each class using a Gaussian Mixture Model (GMM) with 128 mixtures. We use diagonal covariance for the mixtures and choose the best fit among three random starts. To classify an unknown test recording, we again extract features for non overlapping windows of 1s duration and the class with the highest overall posterior likelihood is chosen as the label.

2.3. Adaptation

We refer to adaptation as the changes in the system that take place upon a given task. In the auditory system, top-down attention mechanisms modulate the gain of neurons at the sensory representation stage [11, 12], and are also known to operate at the object representation stage[15, 17, 18].

2.3.1. Sensory Adaptation

We implement sensory adaptation similar to a Bayesian framework where the class posterior is modulated by the both the sensory mapping as well as priors about the class,

representing the general knowledge about the attended class; here captured by $\bar{\mathfrak{R}}(f, \mathfrak{s}, \mathfrak{r})$, the average sensory representation for the target class across the training data. Here, it is applied in a multiplicative fashion, affecting the gains of the modulation filters, as given in Eq. 4. α controls the degree to which the representation is changed, and can be varied between 0 (no change) and 1 (maximum change).

$$\mathfrak{R}(f, t; \mathfrak{s}, \mathfrak{r}) = \left(1 - \alpha + \alpha \frac{\bar{\mathfrak{R}}(.)}{\max \mathfrak{R}}\right) \times |y(f, t) \otimes MF| \quad (4)$$

2.3.2. Object Adaptation

To adapt the object representation stage we assume we have some training examples drawn from the current scenario (i.e. same target, same signal to noise ratio etc.). We then use these examples X to adapt the trained GMM of the target class to the new condition. This is done using the MAP adaptation technique which has been proven useful for speaker verification [23], image segmentation [24], EEG verification [25], etc. The new model parameters $\hat{\theta}$ are chosen as in Eq. 5.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta)^{1-\gamma} \cdot p(\theta)^{\gamma} \quad (5)$$

where $\gamma = (1 + r)^{-1}$ and r is the relevance parameter which controls the amount of adaptation. Increasing values of r leads to more reliance of the new data. We adapt only the means and the probabilities of each mixture. The specifics of MAP adaptation can be found in [23].

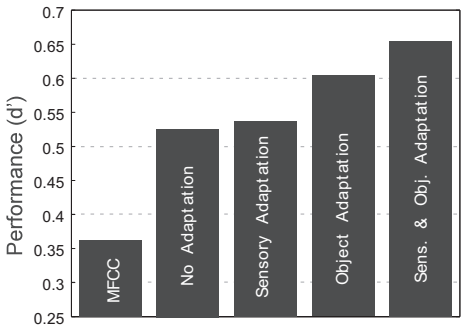

Fig. 2. Performance (d') of the baseline MFCC and proposed model with and without attention mechanism.

2.4. Experimental Setup

The task of auditory scene classification is done on the BBC Sound Effects Database [26]. We chose 12 scene classes containing 1954 recordings amounting to 46 hours of data. These wavefiles are first downsampled to 16 kHz and pre-emphasized with filter coefficients [1 -0.97]. The recordings are then randomly divided into training and test according to a 9:1 ratio. To simulate a multisource scenario, we mix each scene of interest with other recordings randomly chosen from a different scene class, at varying target to masker ratios (TMR) ranging from -20 dB to 20 dB in steps of 5 dB. Furthermore, we generate a set of priors for each scene at each TMR level to be used for object adaptation (Sec. 2.3.2), which consists of 180 randomly chosen 1 second segments from train data mixed with other scene classes.

Performance is measured using the dprime(d') metric. It is defined as d'=Z(hit rate)-Z(false alarm rate) where Z is the inverse cumulative distribution function of a standard normal distribution. We calculate d' for each target class at each considered TMR and report the overall average. This measure has the advantage over classification accuracy of incorporating not only the hit rate, but also false alarm rate.

The proposed system is compared to a system where MFCC representation [2, 8] is used instead of the sensory representation stage. 13 dimensional MFCC coefficients are calculated using a Hamming window of length 25ms with an overlap of 15 ms. The C0 energy coefficient is ignored as our analysis suggested that it is not useful. The mean, standard deviation and skew of these 12 coefficients is calculated over the duration of the segment considered and concatenated resulting in a 36 dimensional feature representation.

3. RESULTS

The proposed system is designed to attend to a target class in a mixed class scenario. When the sensory adaptation parameter α is set to 0 and object adaptation parameter r is set to ∞, the system is denied any adaptation to the task. The performance of such a system is d' =0.53. When MFCC features are used instead, the performance is 0.36 (Fig. 2). This relative improvement of 45% shows that the sensory representation by itself is able to better capture relevant characteristics of individual scene classes.

Next, we test the system with only sensory adaptation by varying α over a range of values and setting $r = \infty$. When the task is to attend to a particular target, we adapt the sensory representation using prior knowledge of the target class as explained in Sec. 2.3.1. This system is tested against the entire test set at each TMR value. This results in a classification confusion matrix for each TMR and target class. The average d' over all TMRs and target classes is considered, which yields a small improvement in performance to 0.54 or 49% relative when $\alpha = 0.2$. This is consistent with physiological studies which show the effect of task related attention on the gain of neurons in terms of α to be in the range of 0.1 to 0.35 [27]. Similarly, we also test the performance of the system with only object adaptation by varying r and setting $\alpha = 0$. In this case, given a particular target, we adapt the object representation stage to that target class as explained in Sec. 2.3.2. This system shows a marked improvement in d' to 0.6 (See Fig. 2) when $r = 0.13$ which is a relative improvement of 67% . This suggests that object adaptation is more effective in enhancing the performance of the system as compared to sensory adaptation. This is not surprising as we assume the additional knowledge of TMR during object adaptation.

We then consider the situation where both sensory and

r	α										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.03	62	78	71	62	51	79	38	21	1	-31	-53
0.06	64	63	77	74	43	77	32	27	-4	-36	-69
0.13	67	70	70	**81**	56	58	50	9	8	-42	-73
0.25	61	64	71	79	56	29	40	39	7	-34	-60
0.5	62	55	53	40	32	36	20	50	32	-0	-47
1	47	50	62	44	2	-1	4	39	32	23	-3
2	46	49	56	43	-22	-26	-9	-6	21	17	-1
∞	45	46	49	30	-2	-24	-32	-45	-37	-50	-53

Table 1. Relative improvement (%) in performance, compared to MFCC, as a function of the adaptation parameters r and α.

object adaptation are applied. We vary both α and r over a range of values and the results are shown in Table 1. The best performance of 0.65 is obtained when $\alpha = 0.3$ and $r = 0.13$. This relative improvement of 81% in performance shows that the two different adaptation techniques can indeed be applied simultaneously to improve the performance further. The choice of α is again consistent with physiological findings.

A comparison of performance for various classes, between the system without any adaptation and with both sensory and object adaptation, as shown in Table 2, reveals interesting traits. Some classes like Foley, Sports and Water are helped by the adaptation mechanisms, while the performance on some classes like Emergency, Household and Weather is deterred. This could be due to the fact that classes like Weather and Household are ambiguous classes, for example, Weather could contain wind sounds and water sounds (rain) and Household could also contain a wide variety of sounds present in a house including water sounds (flowing tap water).

4. CONCLUSIONS

The auditory system maps the acoustic signal into a high-dimensional, redundant mapping which highlights all the relevant spectral and temporal cues in an auditory object. This rich space allows for tracking sound events along multiple time constants, thus providing the ability to selectively enhance or ignore different components of the acoustic environment. In contrast, compact representations that many audio processing approaches strive for (e.g. MFCCs) rather emphasize tight, statistically independent and reduced features that are amenable as front-ends to classification techniques. The current work shows that this high-dimensional mapping outperforms MFCCs by a relative improvement of 45%. Moreover, such high-dimensional representation facilitates incorporating attentional control, which highlights relevant information, segregates pertinent auditory objects and provides a framework to integrate prior knowledge. In this work, attentional feedback is implemented both at the sensory stage (performance gains of 49%); as well as at the object decision stage (performance gains of 67%). Both mechanisms work synergetically. When both are applied, we observe relative improvement of 81% over the baseline system. Neurophysiological evidence does indeed support a role of attentional feedback *both* at the feature-level as well as the object-level [16], with modulatory effects operating at the level of sensory cortex [11, 12, 27] and frontal decision making cortical areas [14, 28].

Class	Model		
	None (d')	Both (d')	Relative Improvement(%)
Emergency	0.7718	0.5189	-33
Foley	0.581	1.4847	156
Humans	0.5501	0.6914	26
Industry	0.5037	0.5652	12
Sports	0.6728	1.6298	142
Transportation	0.5803	0.5189	-11
Water	-0.2152	0.5123	338
Animals	0.8793	0.8042	-9
Household	0.624	-0.0966	-115
SciFi	0.6111	1.0072	65
Technology	0.146	0.3651	150
Weather	0.6035	-0.1415	-123

Table 2. Performance(d') comparison of models, without and with both adaptation mechanisms, for each class.

The incorporation of the rich sensory mapping for representing scenes deviates from prior work in the field of scene classification and audio event detection; which generally gravitated towards more compact short-term features [2, 3, 4]. Nevertheless, it is worth noting that a rich (redundant) feature representation has been employed in more specific sound technologies, such as automatic speech recognition, speaker identification and music instrument classification [29, 30] with notable success. Moreover, the use of top-down attention has been limited to data-driven training of classifiers and back-end generative and discriminative systems [23]; and has seldom been incorporated in manipulating the feature representation; except in few systems [31], but not in scene classification. In contrast, the field of visual scene analysis has had much more success in integrating attentional control and prior knowledge with processing of visual scenes and images [32].

5. REFERENCES

[1] T.D. Griffiths and J.D. Warren, "What is an auditory object?," *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 887–892, 11 2004.

[2] O. Kalini, S. Sundaram, and S. Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP), Rio de Janeiro, Brazil*, October 5-7, 2009.

[3] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *Proceedings of ICAASP'09*, april 2009, pp. 1973 –1976.

[4] X. Zhuang, X. Zhou, T. Huang, and M. Hasegawa-Jhonson, "Feature analysis and selection for acoustic event detection," in *Proceedings of ICAASP08*, 2008.

[5] A. Temko and C. Nadeu, "Classification of acoustic events using svm-based clustering schemes," *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[6] R. Cai, L. Lu, A. Hanjalic, H. Zhang, and L. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 14, no. 3, pp. 1026–1039, May 2006.

[7] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, New York, NY, USA, 2008, MIR '08, pp. 105–112, ACM.

[8] K. Patil and M. Elhilali, "Goal-oriented auditory scene recognition," in *Proceedings of INTERSPEECH 2012*, Portland, USA, September 2012.

[9] L. Miller, M. Escabi, H. Read, and C. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex.," *J Neurophysiol*, vol. 87, no. 1, pp. 516–527, Jan 2002.

[10] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention–focusing the searchlight on sound," vol. 17, no. 4, pp. 437–455, 2007, AN: 17714933; J2: Current opinion in neurobiology; R01 DC005779/DC/United States NIDCD Journal Article Research Support, N.I.H., Extramural Review England; ID: 177.

[11] C.M. Karns and R.T. Knight, "Intermodal auditory, visual, and tactile attention modulates early stages of neural processing," *Journal of Cognitive Neuroscience*, vol. 21, pp. 669–683, 2008.

[12] V. Poghosyan and A.A. Ioannides, "Attention modulates earliest responses in the primary auditory and visual cortices," *Neuron*, vol. 58, pp. 802–813, 2008.

[13] L. Whiteley and M. Sahani, "Attention in a bayesian framework," *Frontiers in Human Neuroscience*, vol. 6, pp. 1–21, June 2012.

[14] J.B. Fritz, S.V. David, S. Radtke-Schuller, P. Yin, and S.A. Shamma, "Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex.," *Nat Neurosci*, vol. 13, no. 8, pp. 1011–1019, Aug 2010.

[15] A. Bregman, *Auditory scene analysis: the perceptual organization of sound*, MIT Press, 1990.

[16] K. Krumbholz, S. B. Eickhoff, and G. R. Fink, "Feature- and object-based attentional modulation in the human auditory "where" pathway," *Journal of cognitive neuroscience*, vol. 19, no. 10, pp. 1721–1733, Oct 2007, JID: 8910747; ppublish.

[17] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[18] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Frontiers in Bioscience*, vol. 5, pp. 202–212, 2000.

[19] X. Yang, K. Wang, and S.A. Shamma, "Auditory representations of acoustic signals," *IEEE transactions on information theory*, vol. 38(2), pp. 824–839, March 1992.

[20] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro temporal analysis of speech using 2d gabor filters," in *INTERSPEECH-2007*, 2007, pp. 506–509.

[21] F. E. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *The Journal of Neuroscience*, vol. 20, no. 6, pp. 2315–2331, March 2000.

[22] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21(4), pp. 1253–1278, 2000.

[23] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[24] M. Barnard and J.-M. Odobez, "Robust playfield segmentation using map adaptation," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, aug. 2004, vol. 3, pp. 610 – 613 Vol.3.

[25] S. Marcel and J.D.R. Millan, "Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 743 –752, april 2007.

[26] *The BBC Sound Effects Library Original Series*, http://www.sound-ideas.com, May 2006.

[27] S. Atiani, M. Elhilali, S.V. David, J.B. Fritz, and S.A. Shamma, "Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields.," *Neuron*, vol. 61, no. 3, pp. 467–480, Feb 2009.

[28] R. Westerhausen, M. Moosmann, K. Alho, S-.O. Belsby, H. Hamalainen, S. Medvedev, K. Specht, and K. Hugdahl, "Identification of attention and cognitive control networks in a parametric auditory fmri study," *Neuropsychologia*, vol. 48, no. 7, pp. 2075 – 2081, 2010.

[29] S. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1, 2012.

[30] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: The biological bases of musical timbre perception," *PLoS Comput. Biol.*, vol. 8, no. 11, pp. e1002759, 11 2012.

[31] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," in *Proceedings of INTERSPEECH 2012*, Portland, USA, September 2012.

[32] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vision research*, vol. 50, no. 22, pp. 2295–2307, 2010.