EXPLORING THE ROLE OF TEMPORAL DYNAMICS IN ACOUSTIC SCENE CLASSIFICATION

Debmalya Chakrabarty *Mounya Elhilali*

Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD, USA.

ABSTRACT

Identification of acoustic scenes often relies on finding the most informative features that best characterize the physical nature of sound sources in the scene. In this paper, we propose a framework that provides a detailed local analysis of spectro-temporal modulations augmented with generative modeling that map both the average modulation statistics of the scene using Gaussian Mixture Modeling (GMM) as well temporal trajectories of these modulations using Hidden Markov Modeling (HMM). Our analysis shows that a hybrid system of these two representations can capture the non-trivial commonalities within a sound class and differences between sound classes. The proposed hybrid system outperforms current systems in the literature by about 30 % and surpasses the performance of the individual GMM and HMM systems suggesting that these representations provide complimentary information about acoustic scenes.

Index Terms— Auditory scenes, Specto-temporal modulations, Temporal trajectories, Gussian Mixture Models, Hidden Markov Models

1. INTRODUCTION

Our surrounding soundscapes are constantly changing as we go about our lives; walking from an office to the street to a cafe and carrying conversations along the way. Humans exhibit a great ability at navigating these complex acoustic environments, and can effortlessly parse and identify their acoustic surroundings; in a process called *auditory scene analysis* [1]. This phenomenon describes complex neural and cognitive processes that underly our ability to detect, identify and classify sound objects in complex acoustic environments. Much like one can identify different visual scenes by the attributes of their constituting objects, a similar process takes place allowing our brain to distinguish a human voice from a bird chirp or a car horn [2]. This ability can provide a great degree of robustness and flexibility to technologies like communication aids for sensory-impaired, surveillance and security systems, context aware computing, audio annotation etc.

The 'identity' of an acoustic scene is largely determined by the acoustic characteristics of the sound sources present at the scene. These sources adapt the spectral profile of the signal to reflect the shape and structure of the vibrating bodies, along with trajectories and reflection paths traveled by sounds until they reach the listener's ear or recording device. The analysis of these characteristics for purposes of automatic identification or classification of acoustic scenes has to take into account all the spectral and temporal attributes of the signal. It has to also be sensitive enough to the natural variability in each class of scenes while discriminative enough across classes. A number of scene classification studies have explored the relevance of low-level features in capturing scene characteristics. These features include low-level time based and frequency based descriptors like short-time energy (STE), zero-crossing rate (ZCR), voicing features like periodicity and pitch information, linear predictive coding coefficients (LPC), as well as the energy distribution entropy of discrete Fourier transform components [3, 4, 5, 6, 7, 8]. These reports suggest that low-level acoustic features are powerful in distinguishing simple scenes. In addition, Mel Frequency Cepstral Coefficients (MFCC) have been a popular feature of choice in studies of acoustic scene classification as they are quite powerful in capturing the overall 'transfer function' (or spectral shaping function) of each scene, and have indeed led to a number of successful implementations of event classification systems [9, 10, 11]. However, in case of complex acoustic scenes, the intricate details of the spectral profile and temporal dynamics of sound events in a scene makes applicability of average features rather limited. Use of global representations of a scene such as cepstral coefficients are generally not capable of capturing fine and subtle changes in the spectrum as it evolves over time; especially in case of dynamic and nonstationary scenes. Instead, it is imperative to consider signal features that capture the spectral and temporal modulations (i.e. changes) in the scene over a wide range of resolutions. Gabor features offer such flexibility in time and frequency by tracking the localized spectral and temporal signal changes over various scales [12].

Use of representative features is intricately linked with choice of backend classifiers that are flexible enough to capture variability across scene classes yet stable enough to work with nuances emerging from the signal features. Commonly used learning techniques include K-NN classifiers and Gaussian mixture models (GMM) which have been used to classify auditory scenes into predefined semantic categories [5]. Statistical models like support vector machines (SVM) and Bayesian network (BN) have also been employed to learn the relationships between audio effects and high-level scene representations [11, 13]. Additional techniques employ descriptive statistics of low level acoustic features and quantify their statistical distributions in terms of mean, variance, skewness and kurtosis [3, 7, 14]. More recently, researchers have focused on modeling the mean statistics obtained from spectro-temporal modulation features via discriminative classifier using multilayer perceptrons and have shown that these representations greatly outperform low-level features like MFCC and its statistics in auditory scene classification task [15].

That being said, one of the challenges of the scene classification task is the inherent complexity of describing what a 'scene' is and the degree granularity that is defined with a chosen dataset for analysis. In the commonly used BBC sound effects dataset [16], the sound class labeled *humor* is a more generic class that encompasses instances of individuals cheering or laughing. These two 'subclasses' can be

978-1-4799-7450-4/15/$31.00 ©2015 IEEE

rather heterogeneous in their signal characteristics making the use of average feature profiles rather limited. Instead, it appears that combining a rich representation of spectro-temporal changes in the signal along with their temporal trajectories could provide added flexibility to capture the heterogeneity of the audio samples in each class [17]. In the current work, we explore the use of a hybrid systems that combines use of spectro- temporal modulation features along with their temporal dynamics to represent sound classes. We explore use of temporal trajectories beyond the classical derivative parameters (Δ, $\Delta\Delta$) by using Hidden Markov Modeling (HMM) applied to modulation features.

The organization of this paper is as follows: In section 2, a brief description of the spectro-temporal modulation features used in the proposed system is provided. Section 3 outlines the classification system modeling both mean statistics as well as temporal trajectories of the modulation features. Section 4 describes the experimental set up and scene classification results; while section 5 provides conclusions and discussion of the results.

2. MODULATION BASED FEATURES

The analysis of modulation features in the acoustic signal is performed in two stages. First, a time-frequency auditory spectrogram is extracted based on a model of peripheral processing in the mammalian auditory system [18]. This first stage starts with a bank of 128 asymmetric filters equally-spaced on a logarithmic axis over 5.3 octaves spanning the range 180 Hz to 8000 Hz. Next, the signal undergoes spectral sharpening via first order derivative along the frequency axis followed by half wave rectification and short term integration with $u(t,\tau) = e^{-t/\tau}u(t)$ where $\tau = 2$ ms. This filterbank analysis results in a time-frequency auditory spectrogram represented by $y(t, f)$. The second stage follows to extract modulation features in the signal. This analysis is performed using a bank of two-dimensional Gabor Filters (GF). Each Gabor filter $GF(f, t; \mathbf{s}, \mathbf{r})$ is parameterized by its spectral modulation tuning or scale (**s** in cycles/octave) and temporal modulation tuning or rate (**r** in Hertz). It effectively filters the detailed fluctuations (called modulations) in the spectral and temporal patterns of the signal. The analysis yields a four-dimensional tensor $\mathcal{R}$ parameterized by time **t**, frequency **f**, scale **s** and rate **r** represented as:

$$\mathcal{R}(t, f; \mathbf{s}, \mathbf{r}) = |y(t, f) \otimes_{f,t} GF(f, t; \mathbf{s}, \mathbf{r})| \tag{1}$$

where $\otimes_{f,t}$ denote convolution in time and frequency. The tensor $\mathcal{R}$ is a multi-resolution mapping of the acoustic signal onto a high-dimensional space [19]. This mapping is akin of the rich representation of sounds in the central mammalian auditory system where specro-temporal response fields of cortical neurons [20] can be mapped onto a space tiled by these Gabor filters.

3. CLASSIFICATION OF MODULATION FEATURES

We use the modulation features denoted by $\mathcal{R}$ to build statistical models for the scene classification task. Our analysis contrasts two types of models, as described next:

3.1. Modeling mean statistics of Spectro-Temporal Representation

The first approach builds a generative model of the data in each class based on average statistics of the scenes. Average statistics are obtained from the 4D modulation tensors $\mathcal{R}$ by first integrating the features over the duration of audio segment. For all analyses presented here, we segment recordings of all sound classes over non-overlapping 1s windows. For each segment, we get a mean representation along frequency, rate and scale axes denoted by $\bar{\mathcal{R}}(f, \mathbf{s}, \mathbf{r})$ which can be expressed as:

$$\bar{\mathcal{R}}(f, \mathbf{s}, \mathbf{r}) = E[\mathcal{R}(t, f; \mathbf{s}, \mathbf{r})] \tag{2}$$

The tensor $\bar{\mathcal{R}}$ is further projected onto a lower dimensional space using Tensor Singular Value Decomposition (TSVD) [21]. We keep 420 dimensions that maintain 99 % variance in the data; resulting in a lower-dimensional modulation tensor $\bar{\mathcal{R}}$. Given the use of modulation features over time and frequency, this lower dimensional $\bar{\mathcal{R}}(f, \mathbf{s}, \mathbf{r})$ captures average changes in the audio segment and is used as feature vector to build Gaussian Mixture Models (GMM) [22] of each sound class to learn its inherent statistical characteristics.

3.2. Modeling the temporal trajectories of Spectro-Temporal Representation

Alternatively, we consider a second model that exploits the temporal trajectories of the modulation tensor $\mathcal{R}$. In this case, instead of integrating $\mathcal{R}$ over the audio segment, the temporal trajectories of $\mathcal{R}$ across multiple time frames over the duration of the audio segment are modeled. Here, we contrast two approaches to modeling these temporal dynamics. First, we explore the commonly-used derivative features that concatenate the base features with their respective first (Δ) and second derivative ($\Delta\Delta$) components [23]. In this case, the mean, Δ and $\Delta\Delta$ features are computed from each audio segment $\mathcal{R}$ and concatenated to generate 1260 dimensional feature vector for building GMM models. The statistical models based on this feature representation exploit some degree of information contained in the temporal dynamics of modulation features.

Alternatively, we explicitly model the temporal trajectories of $\mathcal{R}$ using a Hidden Markov Model (HMM) framework [4, 24]. Each audio segment of duration 1 second is divided into fixed number of frames of duration t_δ (t_δ= 16 ms) to obtain a time series. Then, HMM models parameterized by π_s, $P(s_{t+1}|s_t)$ and $P(y_t|s_t)$ are built where s_t denotes hidden states and y_t denotes the actual observation emitted by hidden state at time instant t. π_s denotes the prior distribution of states, $P(s_t|s_{t-1})$ denotes the transition probability matrix and $P(y_t|s_t)$ is the distribution of observations emitted by hidden states mainly modeled as a Gaussian. The hidden states used in our HMM set up represent which of the frequency channels are active at a particular time instant and the transition of one state to another state corresponds to how the activity of one frequency channel changes with respect to other channels over time. The parameters π_s, $P(s_t|s_{t-1})$ and $P(y_t|s_t)$ of the HMM are learned using the Baum-Welch (BW) algorithm as described in [25].

3.3. Fusion of GMM-HMM models

We also investigate a hybrid model that combines both mean modulation statistics obtained from the GMM model with the temporal trajectories tracked by the HMM model. Here, the underlying assumption is that both models provide complimentary information that gives an even better representation of intricate changes and dynamics in a sound class, that each model by itself would fail to capture. The proposed hybrid GMM-HMM system operates by combining the GMM and HMM models for each sound class $c_1, c_2, \dots, c_k$ using a logistic regression [26]:

$$C = \underset{c_1, c_2, \dots, c_k}{\arg\max}\; w_{GMM} L_{GMM} + w_{HMM} L_{HMM} \tag{3}$$

where C is the class to which the test sample gets assigned, L_{HMM} and L_{GMM} are the respective normalized likelihood scores obtained using HMM and GMM models against a test sample. The logistic weights are trained using a subset development set from the database.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Data

The scene recognition experiments are performed on entire dataset from the BBC Sound Effects Library [16]. The database has total of 2400 recordings, amounting to 68 hours of data. The recordings are organized into 17 classes, for example Ambience, Animals, Transportation and Musical etc. We resample each of the recordings in the database to 16 KHz and preprocess them through a pre-emphasis filter with coefficients $[1 \ -0.97]$ in order to boost high frequencies. 80 % of recordings are randomly selected from the database and used as training set. The remaining 20 % are divided into test and development sets. This latter set is used to train the logistic regression model for the hybrid system. We run a 7-fold cross validation on the entire dataset and report mean accuracy and standard deviation across runs.

4.2. Baseline Setup

The proposed system is contrasted against a baseline setup using MFCC features along with their derivative Δ and $\Delta\Delta$ components. Such setup is close to that used in [5]. We compute 13 MFCC features for every frame size of 25 ms with 10 ms overlap. The average statistics, first and second order delta components of MFCC features are computed across these time frames over a duration of 1 second and concatenated to form a 39 dimensional vector. These vectors are then used to build GMM models for each sound class.

4.3. Results and Analysis

Table 1 summarizes the scene classification accuracy using our proposed hybrid system as well as other setups. The results compare the modulation features against the standard MFCC features along with their derivatives (Δ, $\Delta\Delta$). The performance of individual GMM and HMM classifiers using modulation features and their delta components are also reported to assess their respective accuracy values.

A number of interesting observations are worth noting. Firstly, the modulation-based features provide a clear advantage over MFCC features in capturing scene characteristics; even with use of derivative components. Secondly, the use of derivative components with modulation features further improve the accuracy of classification suggesting that temporal dynamics captured in the rate modulation analysis do not sufficiently represent broader temporal changes in the signal that can be better modeled using derivative cues. Thirdly, the HMM system is slightly worse than the GMM system with the derivative features indicating that the mean statistics captured by the modulation features and their dynamics are likely capturing key aspects of each scene that are not well modeled by the HMM system. Consequently, the hybrid system does provide noticeable improvement further corroborating the observation that representing the average distribution of the features with sufficient statistics complements the temporal trajectories in best modeling heterogeneity in sound classes in the BBC dataset.

Features	Classification Accuracy (%)
GMM based MFCC + Δ + $\Delta\Delta$	49.8 ± 9.5
GMM based modulation features	64.6 ± 5.8
GMM based modulation features + Δ + $\Delta\Delta$	66.8 ± 5.1
HMM based modulation features	65.3 ± 6.4
GMM-HMM based modulation features	**76.57 ± 4.3**
GMM-HMM based modulation features + Δ + $\Delta\Delta$	**79.1 ± 4.1**

Table 1: Results obtained using different features and modeling approaches on Scene Classification Task. ± indicates the standard deviation across folds.

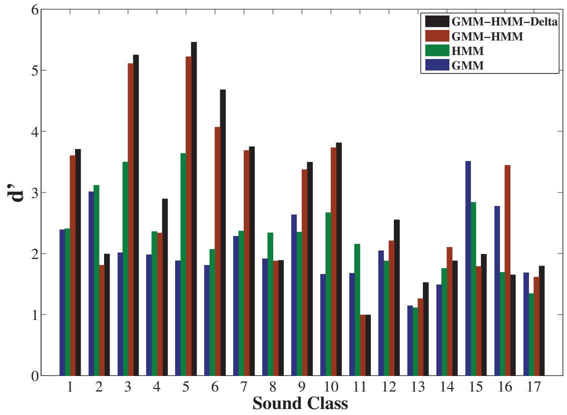

Figure 1: *Classification accuracy of various classifiers in terms of d'.* Sound classes used in the classification task : 1. Ambience 2. Animals 3. Emergency 4. Fire 5. Foley 6. Household 7. Humans 8. Impacts 9. Industry and Machines 10. Musical 11. Science Fiction 12. Sports 13. Technology 14. Transportation 15. Warfare 16. Water 17. Weather

In order to gain a greater insight into the contribution of each of the GMM and HMM models, we examine the performance of these classifiers for each class of scenes using a detection measure of d' [27]. d' is a very popular measure of sensitivity in signal detection theory (SDT) mainly measured in terms of Hit rate (H) corresponding to number of times the model correctly classifies the test signal and False Alarm rate (FA) equal to number of times the model assigns the test signal to wrong class. d' is calculated as : $d' = z(FA) - z(H)$, where z(FA) and z(H) indicate z scores of false alarm and hit rate respectively. Higher value of d' for a class indicates that the model has a high probability of correctly classifying the test signal, hence a model's classification performance can be well represented in terms of its d' value. Figure 1 shows the d' values broken down by class. It is worth noting that most

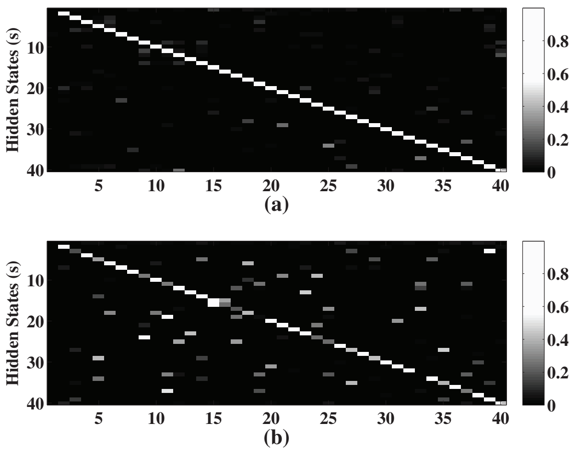

Figure 2: Probability transition matrix ($P(s_t|s_{t-1})$) for class (a) Musical and (b) Water

scenes do exhibit an improved accuracy using the hybrid HMM-GMM system. However, such improvement is not noted across *all* classes. One possible reason for decreased performance of the hybrid system for some sound classes could be due to the greater heterogeneity in those subsets which undermines the score fusion using a simple logistic model. Another interesting observation is the performance of the GMM vs HMM systems across different classes. For instance, the HMM classifier clearly outperforms the GMM for a class like ‘Musical’ which includes different tones with varying degree of spectral and temporal modulations. The temporal characteristics of melodies in this class appear to be best represented using the HMM model; in contrast with a class such as ‘Water’ for instance.

Generally, musical signals have a rich temporal structure and exhibit high degree of temporal regularity [28]. HMM models capture these ‘hidden’ regularities in a much effective manner than GMM by means of its probability transition matrix. Figure 2 depicts the HMM model’s probability transition matrix for 40 states corresponding to classes ‘Musical’ and ‘Water’. In ‘Musical’, there is a very strong activity across the diagonal elements of the transition matrix which means the frequency channels tend to remain in their own state across multiple time frames corresponding to their strong temporal regularity. In case of ‘Water’ , the non zero probabilities in non-diagonal elements of the matrix show that the frequency channels tend to make rapid transitions across each other which affects the temporal structure of the scene. However, because of complimentarity of the information present in temporal structure and average statistics of the scenes, the combination of GMM and HMM models via model fusion gives a tremendous boost in classification accuracy of both the classes as shown in Figure 1.

5. CONCLUSION

In this paper, we examine the role of temporal dynamics of modulation features in capturing intricate details in auditory scenes that extend beyond average statistics of the scene and track the heterogeneous dynamics commonly encountered in these scenes. Specifically, we propose that temporal trajectories of local spectral and temporal profiles do provide complimentary information in addition to their mean statistics. A fusion system based on both representations provides a better model of each sound class relative to the individual models. Such hybrid modeling is crucial in case of complex and unconstrained recordings such as the BBC sound effects data. It is common in such datasets that audio samples representing a similar nominal class but different scenarios are grouped under the same label. Modeling these disparate settings requires not only a representation of the characteristics of the sound sources in the scene, but aspects of their temporal dynamics as well. The proposed model based on a hybrid GMM-HMM model along with derivative components provides noticeable improvement over a MFCC-GMM system (by about 30%) as well as individual GMM or HMM systems (by an average of 14%).

6. REFERENCES

[1] A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound*. Cambridge, Mass.: MIT Press, 1990.

[2] V. T. K. Peltonen, A. J. Eronen, M. P. Parviainen, and A. P. Klapuri, “Recognition of everyday auditory scenes: Potentials, latencies and cues,” in *In Proc. 110th Audio Eng. Soc. Convention*. Hall, 2001.

[3] J. Geiger, B. Schuller, and G. Rigoll, “Large-scale audio feature extraction and SVM for acoustic scene classification,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[4] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, “Audio-based context recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[5] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, “Computational auditory scene recognition,” in *In IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing*, 2001, pp. 1941–1944.

[6] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, “Non-speech audio event detection,” in *Proceedings of ICASSP’09*, April 2008, pp. 1973–1976.

[7] J. Krijnders and G. Holt, “A tone fit feature representation for scene classification,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[8] D. P. Ellis and K. Lee, “Minimal-impact audio-based personal archives,” in *Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, ser. CARPE’04. New York, NY, USA: ACM, 2004, pp. 39–47.

[9] X.Zhuang, X.Zhou, T.Huang, and M.Hasegawa-Johnson, “Feature analysis and selection for acoustic event detection,” in *Proceedings of ICASSP’08*, 2008.

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

[11] W. Nogueira, G. Roma, and P. Herrera, “Sound scene identification based on MFCC, binaural features and a support vector machine (SVM) classifier,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[12] M. Kleinschmidt, “Localized spectro-temporal features for automatic speech recognition,” in *Proceedings of Eurospeech*, 2003, pp. 2573–2576.

[13] L.-H. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, “A flexible framework for key audio effects detection and auditory context inference,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 1026–1039, May 2006.

[14] G. Roma, W. Nogueira, and P. Herrera, “Recurrence quantification analysis features for auditory scene classification,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[15] K. Patil and M. Elhilali, “Goal oriented auditory scene recognition,” in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[16] “The bbc sound effects library original series,” *http://www.soundideas.com*, May 2006.

[17] C. Cotton and D. Ellis, “Spectral vs. spectro-temporal features for acoustic event detection,” in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, Oct 2011, pp. 69–72.

[18] T. Chi, P. Ru, and S. A. Shamma, “Multiresolution spectrotemporal analysis of complex sounds,” *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[19] T. S. Lee, “Image representation using 2d gabor wavelets,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, pp. 959–971, 1996.

[20] R. L. De Valois and K. K. De Valois, “Spatial vision,” *Annual Review of Psychology*, vol. 31, no. 1, pp. 309–341, 1980, pMID: 7362215.

[21] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, “A multilinear singular value decomposition,” *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, March 2000.

[22] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[23] K. Kumar, C. Kim, and R. Stern, “Delta-spectral cepstral coefficients for robust speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4784–4787.

[24] B. Clarkson, N. Sawhney, and A. Pentland, “Auditory context awareness via wearable computing,” in *Proceedings of 1998 Workshop on Perceptual User Interfacces*, 1998.

[25] L. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[26] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2072–2084, Sept 2007.

[27] D. McNicol, *A primer of signal detection theory*. London:George Allen and Unwin, 1972.

[28] E. W. Large and C. Palmer, “Perceiving temporal regularity in music,” *Cognitive Science*, vol. 26, no. 1, pp. 1–37, 2002.