PROBABILISTIC SEGMENTATION AND LABELING OF ETHNOMUSICOLOGICAL FIELD RECORDINGS

Matija Marolt
Faculty of Computer and Information Science
University of Ljubljana, Slovenia
matija.marolt@fri.uni-lj.si

ABSTRACT

The paper presents a method for segmentation and labeling of ethnomusicological field recordings. Field recordings are integral documents of folk music performances and typically contain interviews with performers intertwined with actual performances. As these are live recordings of amateur folk musicians, they may contain interruptions, false starts, environmental noises or other interfering factors. Our goal was to design a robust algorithm that would approximate manual segmentation of field recordings. First, short audio fragments are classified into one of the following categories: speech, solo singing, choir singing, instrumental or bell chiming performance. Then, a set of candidate segment boundaries is obtained by observing how the energy of the signal and its content change, and finally the recording is segmented with a probabilistic model that maximizes the posterior probability of segments given a set of candidate segment boundaries with their probabilities and prior knowledge of lengths of segments belonging to different categories. Evaluation of the algorithm on a set of field recordings from the Ehtnomuse archive is presented.

1. INTRODUCTION

Ethnomusicological field recordings are recordings made "in the field", capturing music in its natural habitat. Starting in the early 20th century and continuing to the present day, ethnomusicologists and folklorists have travelled and made recordings in various parts of the world primarily to preserve folk music, but also to make it available for further researches, such as studies of acculturation and change in music, comparative studies of music cultures and studies of the music making process and its effect through performance. Segmentation of field recordings into meaningful units, such as speech, sung or instrumental parts is one of the first tasks researchers face when a recording is first being studied. It is also a prerequisite for further automatic processing, such as extraction of keywords, melodies and other semantic descriptors.

Segmentation of audio recordings has been extensively explored for applications such as speech recognition (removal of non-speech parts, speaker change detection), segmentation in broadcast news or broadcast monitoring. Typically, the distinction is made between speech, music and silence regions. Approaches to segmentation include either first classifying short periods of the signal into desired classes using some set of features and then making the segmentation [1-3], or first finding change points in features and forming segments and later classifying the segments [4-6]. Authors use a variety of features, classifiers and distances depending on the nature of signals to be segmented. More recently, Ajmera [7] performed classification and segmentation jointly by using a combination of standard hidden Markov models and multilayer perceptrons for speech/music discrimination of broadcast news. Pikrakis et al. [8] used a three step approach: first they identified regions in the signal which are very likely to contain speech or music with a region growing algorithm. Then, they segmented the remaining short (few seconds long) regions with a maximum likelihood model that maximized the probability of class labels given frame-level features and segment length limits. A Bayesian network was used to estimate the posterior probability of a music/speech class label given a set of features. Finally, a boundary correction algorithm was applied to improve the found boundaries. Their use of a probabilistic model is somewhat similar to the proposed segmentation method, but as we describe further on, we use a maximum likelihood approach to segment an entire field recording by first labeling signal fragments, then finding candidate boundaries, and finally maximizing the probability of segmentation considering probabilities of boundaries and segment lengths given their class.

The algorithm presented in this paper was designed to robustly label and segment ethnomusicological field recordings into consistent units, such as speech, sung and instrumental parts. Resulting segmentations should be comparable to manual segmentations researchers make when studying recordings. Field recordings are documents of entire recording sessions and typically contain interviews with performers intertwined with actual performances. As these are live recordings of amateur folk musicians, they usually contain lots of "noise" and interruptions, such as silence when performers momentarily

forget parts of songs, false starts and restarts, laughter, dancing noises, interruptions by other persons, dogs barking or cars driving by. Performances may also change character; singing may become reciting, a second voice may join or drop out of a performance etc.

The described nature of field recordings calls for a robust segmentation algorithm that would not over-segment a recording at each interruption – for example; we are not interested in each boundary separating speech and sung parts, as only some of them are actual segment boundaries. We would also like to distinguish between several different classes of segments and would like to take some prior knowledge of the classes into account. And last, we are not interested in millisecond-exact segment boundaries or exact labeling of each small recording fragment; sometimes placing a boundary between two performances is a very soft decision and accuracy of a few seconds is good enough. Taking these points into account, we propose a three step algorithm for segmentation. First, a standard classification algorithm is used to classify short audio segments into a set of predefined classes. Then, a set of candidate segment boundaries is obtained by observing how the energy and class distribution change, and finally the recording is segmented with a probabilistic model that maximizes the posterior probability of segments given a set of candidate segment boundaries with their probabilities and prior knowledge of lengths of segments belonging to different classes.

2. CLASSIFICATION

Classification of short field recording fragments into a set of predefined categories represents the first part of our segmentation algorithm. We base our work on field recordings from the EthnoMuse digital archive [9]. The archive contains folk song, music and dance collections of the Institute of Ethnomusicology, Scientific Research Centre of Slovene Academy of Sciences and Arts. Audio recordings represent the largest part of the archive and comprise recordings of folk songs and melodies, with the oldest on wax cylinders from 1914 and around 30.000 field recordings on magnetic tape and digital media dating from 1955 onwards. Only parts of the archive are digitally annotated. Field recordings are typically around an hour long and contain interviews with performers intertwined with performances. The latter include singing (solo or group), reciting, instrumental pieces (a large variety of instruments is used, depending on the region), as well as bell chiming, which is a Slovenian folk tradition of playing rhythmic patterns on church bells. The quality of recordings varies a lot and depends on their age, equipment used, location (inside, outside) and type of event (arranged recording session or recording of a public event).

We identified five categories into which field recording fragments are to be classified: speech, solo singing, choir singing (any performance with two or more voices belongs to this class), instrumental (including instrumental with singing) and bell chiming. We then evaluated a set of features often used for speech/music discrimination and timbre recognition to find the ones most suitable for classification into these categories. The following nine features were selected:

- the quotient of RMS energy variance over the squared mean of RMS energy. RMS energy r is defined as:

$$r = \sqrt{\frac{1}{W}\sum_{i=0}^{W-1} x_i^2}\ , \tag{1}$$

  where x represents the time domain signal and W the window size. The feature describes the amount of signal energy fluctuations and is typically larger for speech than for other types of signals;
- mean spectral entropy, as defined by Pikrakis [10]. The entropy represents the instability of signal energy calculated over a number of spectral sub-bands and is typically low for bell chiming recordings, somewhat higher for music, and high for other signal types. It is calculated as:

$$H = -\sum_{i=0}^{L-1} \frac{X_i}{\sum_{j=0}^{L-1} X_j} \log_2 \frac{X_i}{\sum_{j=0}^{L-1} X_j}\ , \tag{2}$$

  where L represents the number of spectral sub-bands and X_i the energy of the i-th sub-band (see [10] for more details);
- variance of spectral entropy deltas. Deltas are calculated as a linear trend over five consecutive windows;
- variances of the first three MFCC coefficients (omitting the zero-*th*). MFCC coefficients describe the shape of the signal spectrum and are thus very appropriate for our classification task;
- variances of deltas of the first three MFCC coefficients (omitting the zero-*th*). Deltas are calculated as a linear trend over five consecutive windows.

To train and test a classifier, we manually labeled 1760 3 second long field recording fragments from the EthnoMuse digital archive. All features were calculated on signals windowed with a 46ms Hamming window with 23ms overlap. Feature means and variances were calculated over 3 second periods, thus taking approx. 130 feature values into account. A multinomial logistic regression classifier [11] was chosen for classification, because it's simple and gives good results. Furthermore, its output can be regarded as a probability distribution over all classes. We trained the classifier to classify each fragment into one of the five previously described classes. 2/3 of the labeled fragments were used for training and 1/3 for testing. Table 1 shows the average confusion matrix of our classifier for 10 training/test runs. The overall accuracy is at 78% of correctly classified instances.

Most of the errors made by the classification algorithm are easy to explain. The confusion of speech and solo singing segments is understandable, if we take into ac-

count that singers are not professional musicians, they are often old persons and their singing close to reciting or very monotonous. Confusion between solo and choir singing occurs in choir segments sung in unison, as well as duet singing, while instrumental and bell chiming segments are correctly classified in most cases with confusion mostly arising between the two classes.

	classified as				
	speech	solo	choir	instr.	bell ch.
speech	79%	14%	4%	3%	0%
solo singing	13%	61%	24%	1%	1%
choir singing	2%	10%	82%	3%	3%
instrumental	1%	3%	3%	82%	11%
bell chiming	0%	0%	2%	7%	91%

Table 1. Confusion matrix of the classification algorithm.

3. SEGMENTATION

To segment a recording, we first find a set of candidate segment boundaries and calculate the probability of splitting the recording at each boundary. Segmentation is then performed by maximizing the joint probability of all segments, taking prior knowledge of segment lengths of different signal classes into account.

3.1 Finding and Evaluating Candidate Boundaries

We consider two criteria for boundary placement: a criterion based on change in signal energy, such as when performances are separated by regions of silence, and a criterion based on change in signal content, such as when speech is followed by singing. To observe changes in energy, we calculate RMS energy e of the audio signal; changes in signal content are detected by calculating the symmetric Kullback-Leibler (KL) divergence d [12] between probabilities of signal classes as calculated by the logistic classifier described in section 2. We find a set of candidate segment boundaries $\mathfrak{B}$ by low-pass filtering both measures to obtain their filtered versions e^f and d^f and finding all candidate boundary regions (b_l, b_r) that satisfy:

$$\mathfrak{B} = \left\{ (b_l, b_r) \,|\, \forall t \in [b_l, b_r] : \left[\begin{array}{l} e_t < \max\left(e_t^f E_1, E_0\right) \text{ or} \\ d_t > \max\left(d_t^f + D_1, D_0\right) \end{array} \right] \right\}, \quad (3)$$

where E_0 and E_1 are the global and relative thresholds that determine the selection of energy-based candidate boundary regions and D_0 and D_1 the global and relative thresholds that determine the selection of divergence-based candidate boundary regions (see also Figure 1 for illustration).

Thus, the set of all candidate segment boundaries contains regions of the signal where its energy falls below, or the amount of change in signal content rises above an adaptive threshold. This is illustrated in Figure 1, which displays a 13 minute long field recording excerpt. The overall RMS energy e (in dB) is displayed on top, the symmetric KL divergence d below. Both adaptive thresholds are indicated with a dotted line; regions where the curves fall below (energy) or raise above (KL divergence) the threshold represent candidate segment boundaries. True segment boundaries are indicated in the middle. As shown, the candidate boundary regions correspond well with true boundaries. Many segments are clearly separated by regions of silence, as the energy plot shows. On the other hand, KL divergence is high where signal content changes, such as between speech and instrumental or sung parts.

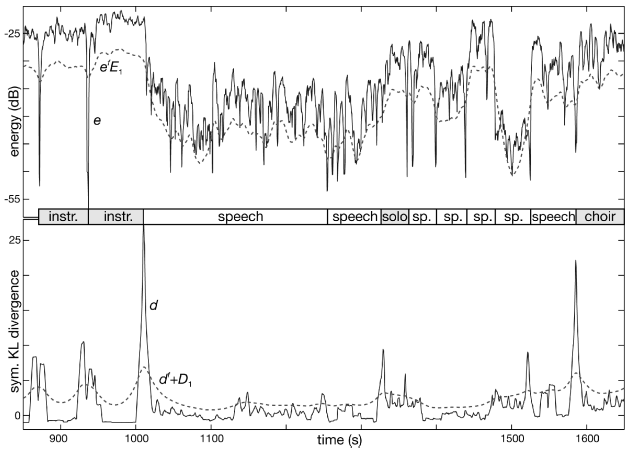

Figure 1. Finding candidate boundaries.

Selecting all of the candidate boundary regions as true boundaries and splitting a recording accordingly is not the best idea; for example energy fluctuates a lot in speech parts (as can be seen in Figure 1) and these parts would consequently be over-segmented. One could attempt to find the best values for relative thresholds D_1 and E_1, but as we show, we can do better by treating the boundary selection process as a classification task. For this purpose, we trained two logistic regression classifiers (one for energy, one for KL divergence) to predict the probability of splitting the segment at a candidate boundary.

The following features were found to be useful for energy-based boundary classification: the amount of signal energy below the energy threshold (s_e) and the maximum difference in signal content to the left and right of the boundary region (m_c). They are calculated as:

$$s_e = \sqrt{\sum_{t=b_l}^{b_r} \max(e_t^f E_1, E_0) - e_t}$$
$$m_c = \frac{1}{N+1} \max_c \left| \sum_{t=b_l-N}^{b_l} P(c_t = c) - \sum_{t=b_r}^{b_r+N} P(c_t = c) \right|, \quad (4)$$

where $P(c_t = c)$ denotes the probability that the signal at time t belongs to class c, as calculated by the classification algorithm presented in section 2 and N the number of frames taken into account to the left or right of the boundary region. The most useful features for the KL diver-

gence-based classifier were found to be the amount of divergence above the threshold (s_d) and the total amount of divergence within the boundary region (t_d):

$$\begin{aligned} s_d &= \log\left(1+\sum_{t=b_l}^{b_r} d_t - \max(d_t^f + D_1, D_0)\right) \\ t_d &= \log\left(1+\sum_{t=b_l}^{b_r} d_t\right) \end{aligned} \quad . \qquad (5)$$

Both classifiers were trained and tested on a set of 30 field recordings from the Ethnomuse archive, which were manually segmented and labeled, containing a total of 840 segments. The classifiers were trained to predict whether a found candidate boundary represents a true segment boundary or not. RMS energy e_t was calculated as the average RMS energy within a 3s window around t and a step size of 0.5s. Symmetric KL divergence d_t was calculated between 10 second long segments to the left and right of t with the same step size. Such large window sizes were chosen primarily to make the algorithm more robust to "noise" in performances, such as false starts, performers forgetting songs, interruptions etc. To obtain the smoothed vectors e^f and d^f, we zero-phase filtered e and d with a first order low-pass Butterworth filter with cutoff frequency of 0.01π. The values of other parameters were experimentally obtained and set to: E_1=0.2, E_0=10^{-6}, D_1=0.1, D_0=3 and N=9. Using these parameters, we extracted approximately 2400 candidate boundary regions from the field recordings and used two thirds of this set to train each classifier to predict whether a candidate boundary is a true segment boundary or not. We evaluated the performance of the two classifiers on the remaining third of the dataset and compared it to an alternative of using an optimal fixed threshold for candidate selection. Table 2 displays average precision and recall scores on the test set for 10 training/test runs. Compared to choosing a fixed threshold for boundary selection, logistic classifiers improve the accuracy of selection. An additional advantage is that their output can be regarded as the probability of splitting the recording at a candidate boundary; a fact exploited by our segmentation algorithm described in section 3.2.

criterion	select. method	precision	recall
energy	best fixed threshold	0.71	0.57
	logistic classifier	0.7	0.67
KL divergence	best fixed threshold	0.77	0.71
	logistic classifier	0.79	0.78

Table 2. Selection of boundary candidates.

3.2 Segmentation algorithm

We perform segmentation by following the logic of Bayesian modeling and infer the most probable segmentation by maximizing:

$$P(seg \mid data) \propto P(data \mid seg)P(seg) \qquad (6)$$

To obtain a generative segmentation model, we define segmentation as a sequence of segments S_{i1}, S_{i2}, ..., S_{iN}, $0<i1<i2< \ldots <i\mathrm{N}$, where S_{i1} starts at time 0 and ends at candidate boundary B_{i1}, S_{i2} starts at candidate boundary B_{i1} and ends at B_{i2}, S_{i3} starts at B_{i2} and ends at B_{i3} and so on. We treat each candidate boundary $B_t \in \mathfrak{B}$ as a discrete random variable with two outcomes: either the candidate boundary represents an actual boundary and splits the recording into two segments, or not. The probability mass function for the variable is defined by outputs of the energy (P_e) and KL divergence (P_{kl}) classifiers, as described in section 3.1:

$$P(B_t = true) = \max\left(P_e(B_t), P_{kl}(B_t)\right) \qquad (7)$$

In our model, the probability of each segment is only dependent on location of the previous segment, so we can express the joint probability of all segments as:

$$P(S_{i1})P(S_{i2} \mid S_{i1})P(S_{i3} \mid S_{i2}) \cdot \ldots \cdot P(S_{iN} \mid S_{iN-1}) . \qquad (8)$$

To calculate the probability of segment S_i given S_j, we must consider all candidate boundaries within the segment, as well as its duration. If the segment is to start at time j and end at i, values of all candidate boundary variables within the segment must be *false*, while the value of candidate boundary variable at time i must be *true*. Segmentation is further constrained by our previous knowledge of typical lengths of segments given their class, leading to the following formulation:

$$P(S_i \mid S_j) = P(D_i \mid S_i, S_j)P(B_i = true)\prod_{j<k<i} P(B_k = false) \quad . \qquad (9)$$

Equation (8) then becomes:

$$\prod_{(i,j)\in\mathfrak{S}} \left(P(B_i = true)P(D_i \mid S_i, S_j)\right) \times \prod_{j\notin\mathfrak{S}} P(B_j = false) \quad , \qquad (10)$$

where $\mathfrak{S}$ is the set of all segment indices and (i,j) a pair of consecutive indices from this set.

Probability of segment duration given its boundaries is dependent on the class of the segment, as calculated by the classifier presented in section 2. By analyzing durations of segments in our collection of field recordings, we estimated the means and standard deviations for all segment classes (μ_c, σ_c); for example the duration of speech segments varies a lot and ranges from several seconds to over ten minutes, while the average length of choir singing segments is around three minutes and their standard deviation below two minutes. By additionally enforcing minimal segment duration D_{min}, we obtain the following expression:

$$P(D_i \mid S_i, S_j) = \begin{cases} \sum_c P(C_i = c \mid S_i, S_j)\, G(i-j, \mu_c, \sigma_c) \\ 0, \; i - j < D_{min} \end{cases} \quad , \qquad (11)$$

where $P(C_i=c|S_i, S_j)$ represents the probability that segment S_i belongs to class c and is calculated as the average

probability of classification of frames within the segment into class *c*. *G* is the unscaled Gaussian function.

To find the sequence of segments that maximizes Equation (10) and thus provides an optimal solution, we resort to dynamic programming that leads us to a simple and efficient solution. For each segment S_i ending at the candidate boundary B_i we can calculate the most probable segmentation that ends with this boundary $d(S_i)$ by the following rules:

$$d(S_i) = \begin{cases} 0.5 & i = 0 \\ P(B_i = true) \max\limits_{0<=j<i} \left(d(S_j)\, c(i,j) \right) & i > 0 \end{cases}, \quad (12)$$

$$c(i,j) = P(D_i \mid S_i, S_j) \prod_{j<k<i} P(B_k = false)$$

where S_0 represents the segment boundary at time 0; S_0 is a boundary if a performance starts at time 0, or not if there is silence or noise present, so we give it a probability of 0.5.

In our implementation, we minimize the negative log-likelihood of segmentation, so all products become summations. When the function $d(S_i)$ is calculated for all candidate boundaries, the most likely segmentation can be recovered by tracking back the calculation and retrieving optimal boundary indices.

After segmentation is calculated, segments can be labeled by finding the class *c* that maximizes $P(C_i=c|S_i, S_j)$; as mentioned before, the latter and is calculated as the average probability of classification of frames within the segment S_i into class *c*.

3.3 Evaluation

As with boundary selection, we evaluated our segmentation algorithm on a set of 30 field recordings from the Ethnomuse archive, which were manually segmented and labeled, containing a total of 840 segments. Because of its specific nature, it is difficult to directly compare the algorithm to other segmentation approaches. We therefore provide a comparison of the proposed method to a simple thresholding algorithm, where segments are formed by thresholding either the energy, KL divergence or the maximal energy/KL divergence candidate boundary probabilities. Results are given in Table 3. Average precision and recall scores of true vs. estimated segment boundaries for all 30 recordings for the three thresholding and the proposed probabilistic method are shown.

	average precision	average recall
thresholding $P_e(B_t)$	0.61	0.61
thresholding $P_d(B_t)$	0.65	0.64
thresholding $\max(P_e(B_t),P_d(B_t))$	0.73	0.78
proposed algorithm	0.78	0.81

Table 3. Comparison of segmentation algorithms.

The probabilistic algorithm is quite robust and improves segmentation accuracy over the more naive thresholding approaches. Most of the false positives occur in speech sections containing very long regions of silence that for example occur when people reflect on past events (consequently causing large drops in energy), or in solo singing performances that are interleaved with reciting or spoken statements, such as "this is repeated three times and we start dancing in a circle so and so ..." (causing high KL divergence). False negatives occur when performances follow each other without significant changes, for example several songs sung in a row almost without interruptions, or when the start or end of a segment is missed, because it interleaves with speech, so that the boundary is placed either too soon or too late in a recording.

To evaluate the influence of the choice of relative and global thresholds (see eq. (3)) on segmentation, we evaluated the algorithm's performance by varying values of the four thresholds individually, with other parameters fixed. The resulting precision/recall curves are given in Figure 2.

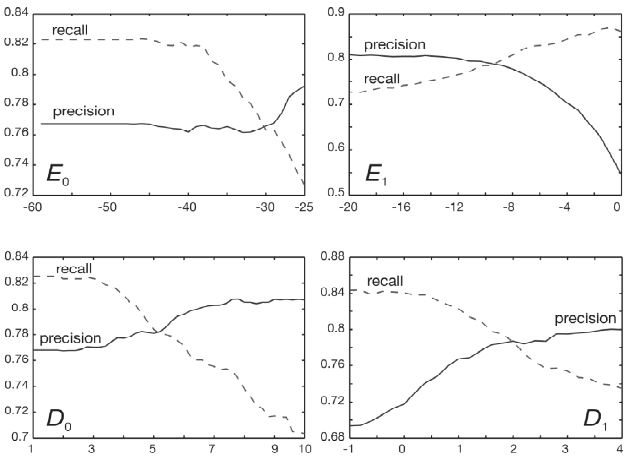

Figure 2. Precision/recall curves obtained by varying the four thresholds that influence candidate boundary region selection: E_0 and E_1 for energy (both are shown in dB), D_0 and D_1 for KL divergence curves.

We can observe that precision is only marginally affected by both global thresholds (E_0 and D_0) – raising them will result in a smaller number of boundaries found, thus decreasing recall, while precision will not increase by much, as the false positives seem to be almost equally spread between weak (low global threshold) and strong (high global threshold) candidate boundary regions. On the other hand, precision is more strongly affected by relative threshold selection (E_1 and D_1); small relative threshold values will result in many false positives, as any significant drop in energy or rise in the KL divergence curve will result in a new boundary candidate. Higher values increase precision and decrease recall, as expected.

The accuracy of classification of correctly found segments into one of the five classes is 86%; errors are similar to the ones described in section 2.

4. CONCLUSION

The proposed algorithm for segmentation and labeling of ethnomusicological field recordings provides a good starting point for further development of automatic methods for analysis of such recordings. Its accuracy is good enough for practical use and the algorithm has already been integrated into the tools of the Ethnomuse archive and is available to its users. For further improvements, we need to start looking into the inner structure of each segment, which may help us to improve the found boundaries. We also plan to explore hierarchical segment classification to classify instrumental segments into typical ensemble types, speech and singing segments into male and female etc.

Acknowledgments. This work was supported in part by the Slovenian Government-Founded R&D project EthnoCatalogue: creating semantic descriptions of Slovene folk song and music.

5. REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 1331-1334 vol.2.

[2] L. Lie*, et al.*, "Content-based audio segmentation using support vector machines," in *IEEE International Conference on Multimedia and Expo,* 2001, pp. 749-752.

[3] G. Williams and D. P. W. Ellis, "Speech/music Discrimination Based On Posterior Probability Features," in *Eurospeech'99*, Budapest, Hungary, 1999, pp. II-687-690.

[4] G. Tzanetakis and P. Cook, "Multifeature audio segmentation for browsing and annotation," in *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, 1999, pp. 103-106.

[5] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *Multimedia, IEEE Transactions on,* vol. 7, pp. 155-166, 2005.

[6] M. Cettolo*, et al.*, "Evaluation of BIC-based algorithms for audio segmentation," *Computer Speech & Language,* vol. 19, pp. 147-170, 2005.

[7] J. Ajmera, "Robust Audio Segmentation," Ph.D., Faculte des sciences et techniques de l'ingenieur, Ecole Polytechnique Federale de Lausanne, Lausanne, 2004.

[8] A. Pikrakis*, et al.*, "A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks," *Multimedia, IEEE Transactions on,* vol. 10, pp. 846-857, 2008.

[9] M. Marolt*, et al.*, "Ethnomuse: Archiving Folk Music and Dance Culture," in *Eurocon 2009*, St. Petersburg, Russia, 2009.

[10] A. Pikrakis*, et al.*, "A computationally efficient speech/music discriminator for radio recordings," in *ISMIR 2006, 7th International Conference on Music Information Retrieval*, Victoria, Canada, 2006.

[11] I. H. Witten and E. Frank, *Data Mining*. San Francisco, USA: Morgan Kaufmann, 2005.

[12] W. D. Penny "Kullback-Liebler divergences of normal, gamma, Dirichlet and Wishart densities," *Technical report*, Wellcome Department of Cognitive Neurology, London, UK, 2001.