LISTENING LEVEL CHANGES MUSIC SIMILARITY

Michael J. Terrell György Fazekas Andrew J. R. Simpson Jordan Smith Simon Dixon
Centre for Digital Music, Queen Mary University of London
Mile End Road, London, E1 4NS, UK
michael.terrell@eecs.qmul.ac.uk, gyorgy.fazekas@eecs.qmul.ac.uk
andy.simpson@eecs.qmul.ac.uk, jordan.smith@eecs.qmul.ac.uk
simon.dixon@eecs.qmul.ac.uk

ABSTRACT

We examine the effect of listening level, i.e. the absolute sound pressure level at which sounds are reproduced, on music similarity, and in particular, on playlist generation. Current methods commonly use similarity metrics based on Mel-frequency cepstral coefficients (MFCCs), which are derived from the objective frequency spectrum of a sound. We follow this approach, but use the level-dependent auditory spectrum, evaluated using the loudness models of Glasberg and Moore, at three listening levels, to produce auditory spectrum cepstral coefficients (ASCCs). The ASCCs are used to generate sets of playlists at each listening level, using a typical method, and these playlists were found to differ greatly. From this we conclude that music recommendation systems could be made more perceptually relevant if listening level information were included. We discuss the findings in relation to other fields within MIR where inclusion of listening level might also be of benefit.

1. INTRODUCTION

The auditory system can be thought of, in signal processing terms, as a level-dependent filter bank, where each component is known as an auditory filter [15]. Incoming sound is first processed by the frequency and direction dependent filter of the pinna (outer ear), before passing through the ear canal, which acts as a narrowband resonant amplifier. The acoustic pressure at the ear-drum is mechanically transmitted, via the amplifying stage of the middle-ear ossicles, to the fluid of the cochlea (inner ear) via the oval window [19]. Due to continuous variation in mass and stiffness along the basilar membrane, the cochlea provides a tonotopic representation (arranged in order of frequency) of sound energy spectrum that is broadly consistent with Fourier analysis.

Within the cochlea, inner hair cells are tonotopically arranged along the basilar membrane. The inner hair cells are innervated with neurons that provide the firing-rate coded signal that is sent to the brain via the auditory nerve. The inner hair cells are accompanied by respective outer hair cells. Pressure gradients in the cochlear fluid cause the inner hair cells at any given location to be deflected in a shearing motion which results from place-frequency dependent resonance of the basilar membrane. At the same time, the motile outer hair cells act in phase-locked synchrony to amplify the excitation. This system is known as the cochlear amplifier.

Each inner hair cell is innervated with a population of neurons that code the local signal in terms of the rate-level function (the function that relates the rate of neuron firing to the perceived intensity level). The stochastic firing rate-level function of a neuron, or a population of neurons, can be thought of as having three distinct stages: spontaneous firing, threshold, and saturation. Below threshold, the neuron fires randomly at a low rate. Between threshold and saturation, the function is close to linear and provides a good coding of level. Above saturation point, increase in level does not result in a proportional increase in firing rate. Thus, with increase in sound pressure level, an increasing area of inner hair cells on the basilar membrane are excited beyond neural threshold. Within the context of the excitation pattern model described above, this is known as spread of excitation.

The action of the cochlear amplifier gives rise to strongly level-dependent tuning of the auditory filter. At low levels, the phase-locked action of outer hair cells provides tonotopically localised amplification, which results in a narrow auditory filter. At high sound pressure levels, the cochlear amplifier is not able to contribute amplification, due to mechanical limits, and so the auditory filter becomes broader with increase in level.

The parameters of the human auditory filter have been determined using psychophysical methods [17] and are represented in terms of equivalent rectangular bandwidth (ERB). Within the music information retrieval (MIR) community, the auditory filters are typically more broadly represented in terms of the approximately analogous Mel frequency scale [21]. The Mel scale is defined in terms of equal pitch distance. Both scales produce a "non-linear mapping" of the frequency domain.

Mel-frequency cepstrum coefficients (MFCC), derived using the discrete cosine transform, have been used for

speech recognition [9], music modelling [13] and music similarity [12]. The ERB scale has been used to improve speech feature extraction [20]. Other related work [9] used a gammatone auditory filter-bank [8] (derived from non-human physiology) in the place of ERBs. The resulting coefficients were referred to as EFCCs.

Thus far, although the MFCCs and EFCCs applied to MIR problems have made some attempt to address the question of perception in terms of frequency warping, no attempt has been made to demonstrate the major level-dependent effects of cochlear processing: (i) absolute threshold, (ii) spread of excitation, (iii) compression, and (iv) masking. In other words, the major parameter of listening level has not been investigated.

At present, MIR is usually based on recordings, which listeners can reproduce at any listening level. Whilst we acknowledge the immediate practical difficulty that this imposes, we believe it is important to determine whether the effects of listening level may be significant. In this article, we use a psychoacoustic model to produce level dependent spectrograms, which incorporate the effects of (i) absolute threshold, (ii) spread of excitation and (iii) compression, and which can be used to evaluate level dependent similarity metrics. The similarity ratings are compared for each listening level to determine whether specific applications of MIR, such as music playlist generation, may be listening level dependent. This article also serves to begin a more general discussion as to the relevance and importance of listening level for other areas within MIR.

2. MODELLING

The loudness models [7, 16] provide a means to predict time and level dependent excitation patterns for time-varying acoustic stimuli. The outer and middle ear stages are modelled as a single FIR filter. Next, a bank of parallel filters is used to calculate spectral magnitude over specific frequency bands. The resulting excitation pattern is then transformed into instantaneous specific loudness (ISL) according to a compressive nonlinearity designed to model the action of the cochlea. The instantaneous specific loudness is essentially a level dependent spectrogram with the frequency axis in the ERB scale. We refer to it as an auditory spectrogram.

We collected a random subset of 500 recordings from the Magnatagatune data set [10]. Magnatagatune is a collection of over 56,000, 30-second music clips from the Magnatune catalogue, with matching tags collected from Law's TagATune game. Our subset of 500 clips has approximately the same proportion of genres as the full data set, including roughly 22% Classical, 17% each of Pop/Rock, Electronic and "Ethnic" or World music, and the rest from assorted genres. The clips, all 44.1kHz, 32kbps mono mp3 files, were obtained using the "Source Only" version of the Magnatagatune data set.

Using the auditory model, auditory spectrograms were estimated at three listening levels for each recording. A 20 ms normalised Hanning window was used with a 50% overlap. The frequency axis was split into ERB bands, which gave 53 discrete frequency bins. The listening levels were characterised by peak sound pressure levels of 40, 80 and 120 dB SPL. The input to the loudness model is a waveform in Pascals (Pa), where a pressure of 1 Pa corresponds to 94 dB SPL. Therefore, in order to convert a normalised digital recording (peak amplitude is 1), s_d, into a pressure signal s_p with a peak level of X SPL, we use,

$$s_p = 10^{\frac{(X-94)}{20}} s_d. \quad (1)$$

Figure 1 shows the auditory spectrograms for a randomly selected recording played at each listening level. At 40 dB SPL it becomes relatively narrow-band due to the high and low frequency energy falling below the absolute thresholds of audibility. At 80 dB the majority of the energy is above absolute threshold and the auditory spectrogram is similar to the objective spectrogram. At 120 dB SPL the spread of excitation causes smearing of the energy across the frequency range, and the recording becomes relatively broadband.

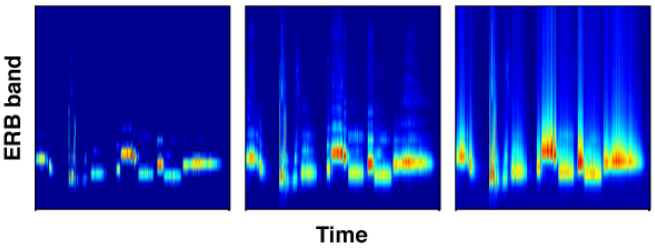

Figure 1. The auditory spectrograms of a randomly selected recodring with peak play-back intensity levels from left to right of: 40, 80 and 120 dB SPL respectively.

3. ANALYSIS

An acoustic model of musical timbre is often a core component of content-based MIR systems. It is fundamental in tasks such as content-based music recommendation [13], playlist generation [18], genre classification [23] and instrument recognition [5]. In our experiments, we choose to follow a deliberately simple, yet widely adopted method of modelling the overall timbre of a recording first by extracting frame-wise cepstral coefficients, and then modelling the overall timbre distribution by fitting a single Gaussian to the resulting coefficient vectors [13]. In order to be able to take the effect of listening level into account, we use a set of auditory spectra cepstral coefficients (termed ASCCs), computed from auditory spectra, calculated using the method outlined in Section 2.

Similarly to MFCCs, the computation of this feature is derived from the computation of the real Cepstrum shown in Equation 2, where $X(\omega)$ represents the Fourier transform of the analysed signal. The cepstrum separates the slowly varying components of a signal from superimposed higher frequency and noise like components. It can be viewed as a rearranged spectrum, such that relatively few coefficients are sufficient to characterise the spectral envelope; however, the higher the number of coefficients, the

more spectral detail is retained.

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(\omega)| \, e^{j\omega n} d\omega \qquad (2)$$

In many applications, including speech recognition and audio similarity analysis, it has become common to characterise short audio segments using a set of cepstral coefficients, such that non-linear frequency warping is used to emphasise perceptually relevant frequencies corresponding to auditory bands. Mel-scaling is the most widely adopted method for this purpose.

Our feature extraction follows a common procedure of computing MFCCs [4]; however instead of using Mel-scaled magnitude spectra, we use auditory spectra estimated at three different listening levels. The auditory spectrograms are logarithmically compressed and then decorrelated using the Discrete Cosine Transform (DCT) given in Equation 3.

$$C(n) = \sum_{i=1}^{M} X(i) \cos \frac{n(i-0.5)\pi}{M}, \text{with } n = 1, 2..., J, \qquad (3)$$

where M is the number of auditory filters, J is the number of ASCCs (typically $J < M$), and $X(i)$ is the log-magnitude output of the i-th filter. These coefficients are then modelled using a single Gaussian characterising the distribution of ASCCs over a song in our collection.

This method makes several simplifying assumptions. For one, it ignores musical structure, and also the fact that the distribution of timbre features is not necessarily Gaussian. A solution to these problems may be the use of Gaussian mixture models (GMM) or a sequence of Gaussians fitted on coherent segments, for instance, a single Gaussian representing each bar or each structural segment of the music, for modelling a track. However, approaches to estimate similarity between these models such as Monte Carlo sampling are computationally expensive. Detailed discussions on timbre models and the effects of the above assumptions can be found, for instance, in [1], [2] and [3]. Besides modelling recordings using a single Gaussian, a further simplifying assumption is introduced by using Gaussians with diagonal covariance. Although modelling timbre using a single Gaussian is a very simple approach, it was shown in [14] that it can perform comparably to mixture models when computing similarity between recorded audio tracks. It was also shown to be effective and computationally efficient for finding similar songs in personal music collections in [11]. An important advantage of using this model is that the similarity between two tracks can be estimated using closed form expressions, such as the Jensen-Shannon (JS) or Kullback-Leibler (KL) divergences. Here, we use the symmetrised KL divergence given in Equation 4, where p and q are Gaussian distributions, with μ mean and Σ covariance, and d is the dimensionality of the feature vectors.

$$\begin{aligned} KL_s(p\|q) &= 2KL(p\|q) + 2KL(q\|p) \\ &= tr(\Sigma_q^{-1}\Sigma_p + \Sigma_p^{-1}\Sigma_q) \\ &\quad +(\mu_p - \mu_q)^T(\Sigma_q^{-1} + \Sigma_p^{-1})(\mu_p - \mu_q) \\ &\quad -2d \end{aligned} \qquad (4)$$

Using this simple model, we calculate symmetric distance matrices holding pair-wise KL-divergences (similarity estimates) between all recordings in our collection. For each distance matrix computation, different sets of ASCCs are used that are calculated from the auditory spectra estimated for different listening levels. The distance matrices are then individually analysed using the methods described in Section 4.1 and 4.2, and the results produced at three different levels are compared.

4. RESULTS

The data set is analysed as per Section 3 to produce a KL divergence rating per pair of recordings at each listening level. To illustrate the approach, 25 tracks from the set (n=500) were selected at random and KL divergence matrices computed at each listening level (40, 80, 120 dB SPL). Figure 2 shows the matrices. Blue indicates low values (similar) and red indicates high values (dissimilar). Figure 3 shows a box-plot of the matrix data. Figs. 2 and 3 clearly illustrate that the similarity ratings are strongly dependent on the listening level. At low level, the set shows a high mean similarity with relatively small variance. At high level the mean similarity is lower and the variance is larger. At the medium level the variance lies between the low and high listening levels.

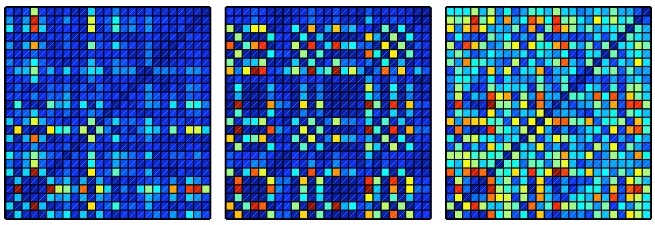

Figure 2. The normalised KL divergence matrices for a subset of recordings with peak intensity levels from left to right of: 40, 80 and 120 dB SPL respectively. Blue indicates low values (similar) and red indicates high values (dissimilar).

Whilst Fig. 2 shows that the similarity ratings are dependent upon the listening level, it is important to determine whether these differences are significant in MIR applications. The application we choose to study is music recommendation. Music recommendation tools generate playlists based on similarity ratings, typically derived from MFCCs. We compared the similarity data across the three intensity levels in two ways: (i) by comparing the ordering of distances within triples, and (ii) by comparing the members of playlists with different seed recordings, and with different playlist sizes.

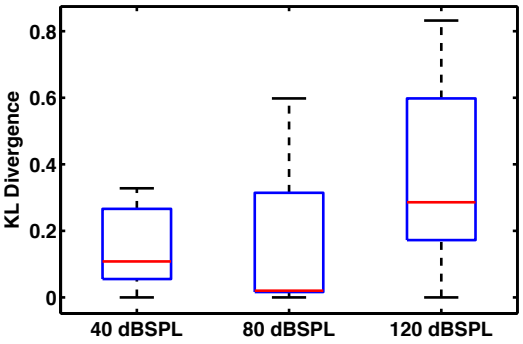

Figure 3. Boxplots of the KL divergence matrices (Fig. 2) at each listening level. Low values correspond to similar recordings and high values to dissimilar recordings.

4.1 Triple analysis

We analysed all subsets of 3 recordings from the dataset and the pair of recordings with minimum distance (in terms of the KL divergence feature space), was identified. The data was compared across listening level, and changes in the closest identified pairs were recorded. For example, if a given triplet (I,J,K) showed that at 40 dB SPL recordings I and J were closest together, but that at 80 dB SPL I and K were closest together, this was recorded as a change. The percentage changes were calculated across all triples and are shown in Table 1. We see around a 30% change in the ordering of triples. This suggests that MIR applications that use similarity metrics, such as playlist generation, will be affected by listening level.

% Change in Triplet Order		
40 vs 80	40 vs 120	80 vs 120
32	29	27

Table 1. The percentage change in the closest identified pair within each set of triples. The column headers refer to the listening levels between which the comparisons were made, i.e. 40 vs 80 relates to comparison of triplet data from the 40 dB SPL and 80 dB SPL sets.

4.2 Playlist generation

Playlists were generated by assigning a seed song, and then identifying the $(n-1)$ closest songs in the similarity space, where n is the size of the playlist. The playlists were compared across listening levels. For example, if a five song playlist is generated for seed song A, where identified songs are (T,U,S,X) at 40 dB SPL, but at 80 dB SPL are (W,T,U,S), the percentage change would be 25%. We do not consider a playlist to have changed if the order of the chosen songs is different.

The mean and 95% confidence intervals are calculated for playlist changes across all seed songs. The mean data are shown in Table 2 using the first 20 ASCCs. Playlist change data using first 12, 20 and 29 ASCCs are plotted in Figure 4. The changes range from 80% for small playlists, to 50% for large playlists.

In order to verify the significance of these changes, an equivalent process is followed but comparisons are made between playlists generated using different numbers of ASCCs at each listening level. These data are shown in Figure 5. The changes range from 50% for small playlists, to 10% for large playlists. For a 10 song playlists, the average change (in the songs added) is: 62% caused by listening level (Fig. 4), and 22% caused by the number of ASCCs used (Fig. 5).

N. Songs	Mean % Change in Playlist Members		
	40 vs 80	40 vs 120	80 vs 120
1	74	67	80
2	69	64	78
3	68	62	76
4	66	59	75
5	66	58	75
6	65	57	74
7	64	56	73
8	63	55	73
9	62	54	72
10	61	53	71
11	61	52	70
12	60	52	70
13	60	51	69
14	59	51	68
15	58	50	68
16	58	49	67
17	58	49	67
18	57	48	66
19	57	47	66
20	57	47	65
21	57	46	65
22	56	46	65
23	56	46	64
24	55	45	64

Table 2. The percentage change in the recommended playlists using the first 20 ASCCs. The column headers refer to: the length of playlist (excluding seed song), $(n-1)$, and the listening levels between which the comparisons were made, i.e. 40 vs 80 relates to comparison of playlists from the 40 dB SPL and 80 dB SPL sets.

5. DISCUSSION

We have demonstrated that the effect of listening level is larger than that of variation of the number of ASCCs used in the playlist generation. The large percentage change in playlist members shown for the comparison between 40-80 dB SPL is perhaps most relevant to the typical MIR end user - such variation in listening levels may be typical in the home (e.g., for radio broadcast). The equally large percentage change shown in the results for the highest sound pressure level (120 dB SPL) may be relevant for the live

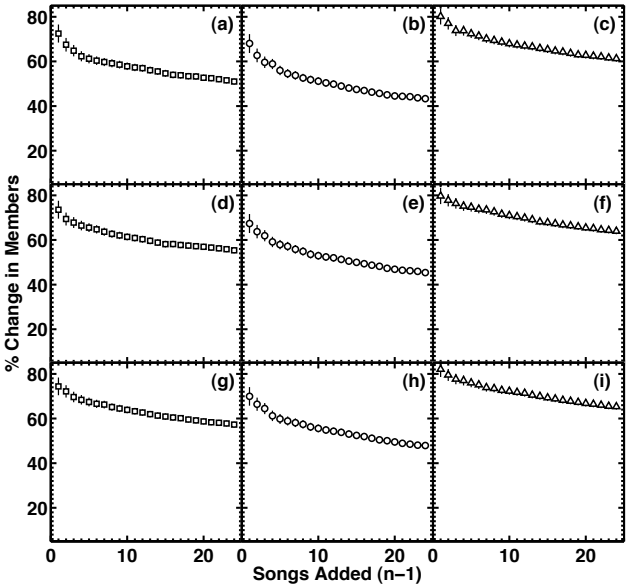

Figure 4. The percentage change in playlist members with listening level as a function of the length of playlist (excluding seed song). The data shown are the mean and 95% confidence intervals across all seed songs. The square, circle and triangle markers show comparisons between: 40 to 80, 40 to 120 and 80 to 120 dB SPL respectively. Figs. (a) to (c) show comparisons using the first 12 ASCCs, (d) to (f) use the first 20, and (g)-(i) use the first 29.

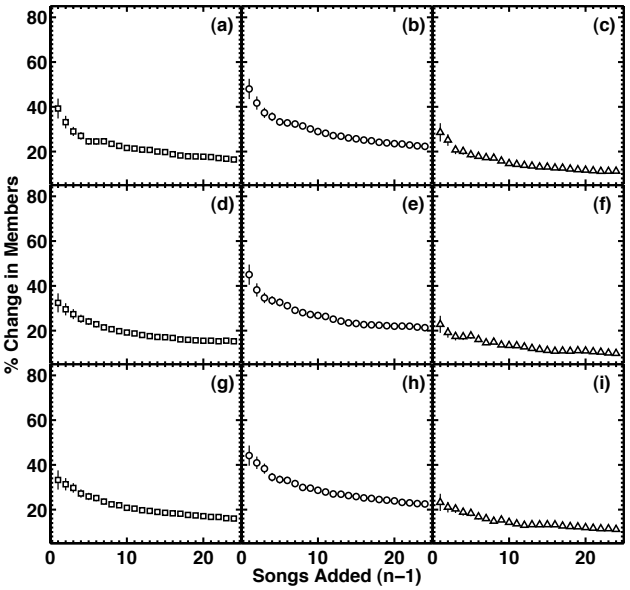

Figure 5. The percentage change in playlist members with the number of ASCCs used as a function of the length of playlist (excluding seed song). The data shown are the mean and 95% confidence intervals across all seed songs. The square, circle and triangle markers show comparisons between: 12 to 20, 12 to 29, and 20 to 29 ASCCs respectively. Figs. (a) to (c) show comparisons at 40 dB SPL, (d) to (f) at 80 dB SPL and (g)-(i) at 120 dB SPL.

sound (or disc jockey) context, where sound levels tend to be higher.

Another conclusion that may be drawn from the analysis is that low listening levels may be considered to produce a homogenization effect by limiting bandwidth (due to absolute thresholds). A similar effect is seen at high levels, where saturation and upward spread of excitation limit the effective number of independent ASCCs. It is conceivable that, given a larger set from which playlist members are drawn, the trends shown in Figs. 4 and 5 would resolve to a more signal or method dependent function, for example, it may be shown that the effect of listening level is more significant on certain genre. Future work should include modelling with larger sets of data.

Although demonstrated here using a music similarity study, the effect of listening level on auditory spectra may have wide ranging implications for MIR theory and practice in general, and initiating this debate was a primary aim of this article. It seems unlikely that changes in listening level will manifest changes in MIR properties relating to musical score (e.g., notation) or structure (e.g., segmentation). However, where MIR methods rely on spectrum (e.g., timbre) some effects of listening level may be expected. For example, speech (or even speaker) recognition in a high noise environment might be enhanced by the proper masking (noise suppression) effects of loud speech in the auditory model. In a more general sense, loudness itself may be a useful perceptual feature for MIR problems. For example, in the creation of a playlist, using a similar procedure to that described in the present paper, loudness and loudness dynamic range may be used to produce a sequence of songs which is tailored for smooth loudness transitions between tracks, and for similar loudness dynamics. Furthermore, incorporation of complete psychoacoustic listening conditions within listening tests designed to validate such perceptual similarity metrics may lead to more meaningful ground truth data.

6. CONCLUSIONS

In this paper we have presented a computational analysis of the effect of listening level on a perceptual music spectrum similarity metric. The similarity matrices and statistical data have shown that the metric is strongly level dependent. The playlist data shows similarly striking effects of listening level. Some general discussion has been given on the immediate implications of the use of listening-level dependent auditory models in MIR and loudness itself has been suggested as possible future similarity feature. The results of this study suggest that more complete data about sound [22] and about music production [6] may be useful to future context specific MIR applications.

7. REFERENCES

[1] J. J. Aucouturier. *Ten experiments on the modelling of polyphonic timbre*. PhD thesis, University of Paris, 2006.

[2] M. Casey, and M. Slaney. The importance of sequences

in musical similarity." *Proc. IEEE Int. Conf. ASSP*, 2006.

[3] M. A. Casey, R. Veltcamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. "Content-based music information retrieval: Current directions and future challenges." *Proceedings of the IEEE*, Vol. 96, No. 4, pp. 668–696, 2008.

[4] S. B. Davis, and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." *Proc. IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366, 1980.

[5] A. Eronen, and A. Klapuri. "Musical instrument recognition using cepstral coefficients and temporal features." *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 753–756, 2000.

[6] G. Fazekas, and M. Sandler. "The Studio Ontology Framework." *Proc. 12th Int. Soc. for Music Information Retrieval, USA.*, 2011.

[7] B. R Glasberg, and B. C. J. Moore. "A model of loudness applicable to time-varying sounds." *J. Audio Eng. Soc*, Vol. 50, pp. 331–342, 2002.

[8] P. I. M Johannesma. "The pre-response stimulus ensemble of neurons in the cochlear nucleus." *Symp. Hear. Theory*, pp. 58–69, 1972.

[9] K. Metha. "Robust front-end and back-end processing for feature extraction for hinhi speech recognition." *Proc. IEEE Int. Conf. ICCIC*, 2010.

[10] E. Law, and L. von Ahn. "Input-agreement: a new mechanism for collecting data using human computation games." *Proc. 27th Int. Conf. on Human Factors in computing systems*, pp. 1197–1206, 2009.

[11] M. Levy, and M. Sandler. "Lightweight measures for timbral similarity of musical audio." *Proc. 1st ACM Workshop on Audio and Music Computing Multimedia*. 2006.

[12] B. Logan, and A. Salomon. "A music similarity function based on signal analysis." *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 745–748, 2001.

[13] B. Logan. "Mel frequency cepstral coefficients for music modeling." *Proc. Int. Symp. on Music Information Retrieval*, 2000.

[14] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis. "Support vector machine active learning for music retrieval." *Proc. Int. Conf. on Music Information Retrieval*, 2005.

[15] B. C. J. Moore. *An Introduction to the Physiology of Hearing*. Academic Press, 1997.

[16] B. C. J. Moore, B. R. Glasberg, and T. Baer. "A model for the prediction of thresholds, loudness, and partial loudness." *J. Audio Eng. Soc*, Vol. 45, pp. 224–240, 1997.

[17] B. C. J. Moore ,and B. R. Glasberg. "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns." *The Journal of the Acoustical Society of America*, Vol. 74, No. 3. pp. 750–753, 1983.

[18] E. Pampalk. *Computational models of music similarity and their application to music information retrieval*. PhD thesis.

[19] J O. Pickles. *An Introduction to the Physiology of Hearing*. 2008.

[20] M. D. Skowronski, and J. G. Harris. "Improving the filter bank of a classic speech feature extraction algorithm." *Proc. Int. Symp. Circuits and Systems*, Vol. 4, pp. 281–284, 2003.

[21] S. S. Stevens, J. Volkmann, and E. B. Newman. "A scale for the measurement of the psychological magnitude pitch." *The Journal of the Acoustical Society of America*, Vol. 8, No. 3, pp. 185–190, 1937.

[22] M. J. Terrell, A. J. R. Simpson, and M. Sandler. "Sounds not signals: A perceptual audio format." *Eng. Brief, AES 132nd Int. Convention*, 2012.

[23] G. Tzanetakis and P. Cook. "Musical genre classification of audio signals." *IEEE Trans. Speech Audio Processing*, Vol. 10, pp. 293–301, 2002.