MODIFIED PERCEPTUAL LINEAR PREDICTION LIFTERED CEPSTRUM (MPLPLC) MODEL FOR POP COVER SONG RECOGNITION

Ning Chen[1] **J. Stephen Downie**[2] **Haidong Xiao**[3]
Yu Zhu[1] **Jie Zhu**[4]
[1] Dept. of Elec. and Comm. Eng., East China Univ. of Sci. and Tech., CHN
[2] Graduate School of Library and Information Science, UIUC, USA
[3] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, CHN
[4] Dept. of Electronic Engineering, Shanghai Jiao Tong University, CHN
nchen@ecust.edu.cn

ABSTRACT

Most of the features of Cover Song Identification (CSI), for example, Pitch Class Profile (PCP) related features, are based on the musical facets shared among cover versions: melody evolution and harmonic progression. In this work, the perceptual feature was studied for CSI. Our idea was to modify the Perceptual Linear Prediction (PLP) model in the field of Automatic Speech Recognition (ASR) by (a) introducing new research achievements in psychophysics, and (b) considering the difference between speech and music signals to make it consistent with human hearing and more suitable for music signal analysis. Furthermore, the obtained Linear Prediction Coefficients (LPCs) were mapped to LPC cepstrum coefficients, on which liftering was applied, to boost the timbre invariance of the resultant feature: Modified Perceptual Linear Prediction Liftered Cepstrum (MPLPLC). Experimental results showed that both LPC cepstrum coefficients mapping and cepstrum liftering were crucial in ensuring the identification power of the MPLPLC feature. The MPLPLC feature outperformed state-of-the-art features in the context of CSI and in resisting instrumental accompaniment variation. This study verifies that the mature techniques in the ASR or Computational Auditory Scene Analysis (CASA) fields may be modified and included to enhance the performance of the Music Information Retrieval (MIR) scheme.

1. INTRODUCTION

Cover Song Identification (CSI) refers to the process of identifying an alternative version, performance, rendition, or recording of a previously recorded musical piece [26]. It has a wide range of applications, such as music collection search and organization, music rights management and licenses, and music creation aids. Inspired by the actual application requirements and researchers' growing interest in identifying near-duplicated versions, CSI has become a dynamic area of study in the Music Information Retrieval (MIR) community over the past decades. As a result, for the first time in 2006, the CSI task was included by the Music Information Retrieval Evaluation eXchange (MIREX), an international community-based framework for the formal evaluation of MIR systems and algorithms [6].

Since there are many different formats of cover version, such as remastering, instrumental, mashup, live performance, acoustic, demo, remix, quotation, medley, and standard, the cover version may differ from the original in timbre, tempo, timing, structure, key, harmonization, lyrics and language, and noise [24]. What remain almost invariable among cover versions are melody evolution and harmonic progression, which form the basis of most existing CSI feature extraction algorithms. Among these features, the Pitch Class Profile (PCP) (or chroma) [9] and related descriptors [3, 7, 19, 25, 26, 31, 33]–which can represent harmonic progression directly–are robust to noise (e.g. ambient noise or percussive sounds) and independent of timbre, played instruments, loudness, and dynamics, have become the most widely-used features for CSI. In [7], the beat-synchronous chroma for two tracks were cross-correlated, from the results of which the sharp peaks indicating good local alignment were looked for to determine the distance between them. This CSI scheme performed the best in the audio CSI task contest of the 2006 MIREX. The Harmonic Pitch Class Profile (HPCP) feature proposed in [12] shared the common properties of PCP, but since it was only based on the peaks of the spectrum within a certain frequency band, it reduced the influence of noisy spectral components. It also took the presence of harmonic frequencies into account and was tuning independent. The CSI scheme based on the HPCP and Q_{max} similarity measure [26,27] achieved the highest identification accuracy in the audio CSI task contest of the 2009 MIREX. In [19], the lower pitch-frequency cepstral coefficients were discarded and the remaining coefficients were projected onto chroma bins to obtain the Chroma DCT-Reduced log Pitch (CRP) feature. The CRP feature achieved high degree of timbre

invariance and, thus, outperformed conventional PCP in the context of music matching and retrieval applications.

We observed that despite the promising achievements of the CSI technique over the last decade, the available CSI schemes cannot perform as well as the human ear does. One possible reason is that the available CSI schemes pay attention solely to the musical facets (e.g. melody evolution and harmonic progression) that are shared among cover versions and do not resemble the way humans process music information at all [24]. In this paper, we propose a perceptually inspired model called the MPLPLC model to process music signals based on the Perceptual Linear Prediction (PLP) model [13] in the ASR field. In the proposed scheme, we will consider equally the various attributes of human auditory processing, the difference between speech and music signals, and the requirements of representing the musical facets shared among cover versions. First, the MPLPLC model uses the Blackman window but not the Hamming window to weight each frame to maintain the harmonic information of the music. Second, it replaces frequency warping on the bark scale with a real filter bank equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale to model the time and frequency resolution of human ears. Third, it substitutes a fixed equal loudness curve for a loudness model suitable for time-varying sounds (speech or music) [11]. Fourth, the hair cell transduction model [17] takes the place of cubic-root intensity-loudness compression to replicate the characteristics of auditory nerve responses, including rectification, compression, spontaneous firing, saturation effects, and adaptation [32]. Last and most important, to make the resulted feature (MPLPLC) suited for the CSI task, the LPCs are transformed into LPC cepstrum coefficients to reduce the correlation between them and their unnecessary sensitivity, the result of which is liftered to achieve some degree of timbre invariance [1, 14].

The identification power and robustness to the variation in instrumental accompaniments of MPLPLC were tested on two different collections. The first was composed of 502 songs and 212 cover sets and the second consisted of 85 cover sets whose cover versions have been performed by the same artist with different instrumental accompaniments. We observed that MPLPLC achieved higher identification accuracy, in terms of the Mean of Average Precision (MAP), the total number of identified covers in the top five (TOP-5), the mean rank of the first identified cover (RANK), and the Mean averaged Reciprocal Rank (MaRR) [23]. It also achieved a higher degree of invariance to instrumental accompaniments than the conventional PLP feature [13] and different PCP-related features: the beat-synchronous chroma [7], the HPCP [12, 26], and the CRP [19]. Experimental results also verified that both the LPC cepstrum coefficients mapping and the cepstrum liftering are crucial in ensuring the identification power of MPLPLC.

The rest of this paper is organized as follows. The signal processing steps involved in the proposed MPLPLC model have been described in detail in Section 2. The performances of the MPLPLC feature in the CSI task in comparison with PLP and other state-of-the-art features have been evaluated and discussed in Section 3. Conclusions and prospects on future work have been given in Section 4.

2. MPLPLC MODEL

A block diagram of the MPLPLC model is shown in Figure 1. The signal processing steps involved in this model are discussed in detail as follows.

2.1 Pre-processing

The input music signal is first converted to mono, 8 kHz and 16 bits per sample version to reduce both the computation time and memory requirements. Then, it is filtered by a preemphasis filter of the form

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

where the coefficient μ is chosen between 0.95 and 0.99. The preemphasis is needed because first, it weakens the influence of low-frequency noise and strengthens the high-frequency signal; second, it reduces the dynamic range of the spectrum to make autoregressive modelling easier [4]; and third, it has been proven helpful in maintaining harmonic information in audio signals [22].

2.2 Enframing

The pre-processed signal is segmented into overlapping frames, denoted as $\{\mathbf{s}_i | i = 1, \cdots, N\}$, and each frame is windowed by the Blackman window [20] to get $\{\mathbf{s}_{w\_i} | i = 1, \cdots, N\}$.

We chose the Blackman window but not the Hamming window because the Blackman window has a wider mainlobe and lower highest side-lobe than the Hamming window [28]. As described in the open course *Audio Signal Processing for Music Applications* [1] , this characteristic of the Blackman window helps to maintain and smooth the peaks in the spectrum corresponding to the harmonics in the music signal.

2.3 Equal Loudness Predicting

To compensate for the frequency-dependent transmission characteristics of the outer ear (pinna and ear canal), the tympanic membrane, and the middle ear (ossicular bones), each windowed frame $\mathbf{s}_{w\_i}$ is filtered by an equal loudness model to simulate the transfer function from the sound field to the oval window of the cochlea [2] to get $\mathbf{s}_{wl\_i}$. In PLP, a fixed equal-loudness curve is combined [13]. However, since a music signal is time-varying and has both short-term loudness (the loudness of a specific note) and long-term loudness (the loudness of a musical phase) [18], the fixed loudness curve is not suited to it. So, Glasberg and Moore's [11] loudness model, which can be applied directly to the sound and works for time-varying sounds, is applied to the MPLPLC model.

[1] https://class.coursera.org/audio-001/lecture/53

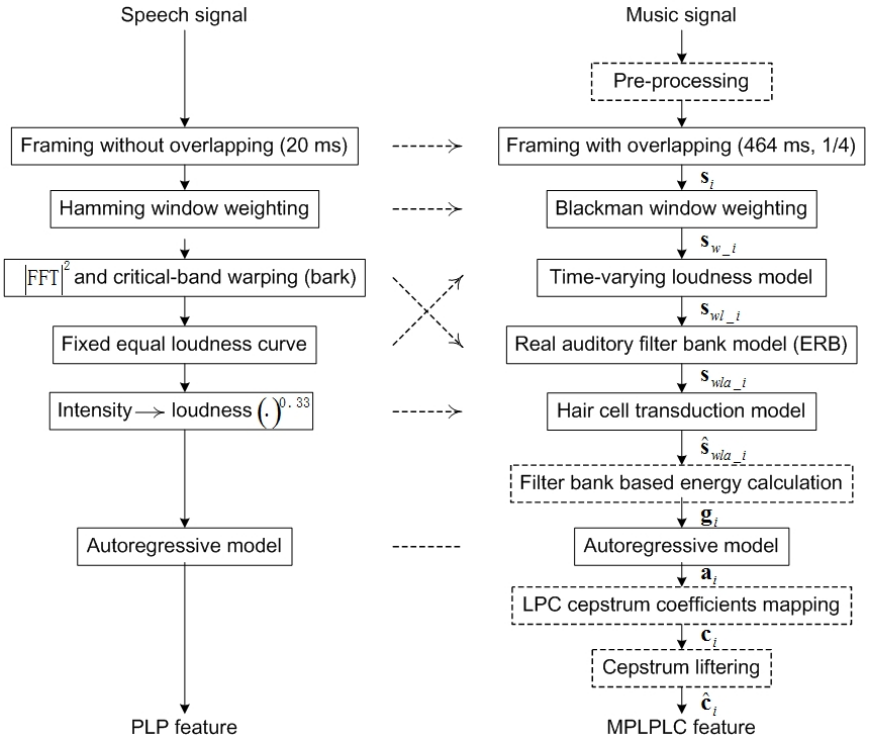


Figure 1. The comparison between the PLP model (left) and MPLPLC model (right).

2.4 Auditory Filter Bank Modeling

To obtain the auditory spectrum, PLP does a critical-band integration after a Fourier Transform (FT) [13]. The problem is that frequency bin in FT is linear, so it has a constant spectral resolution, while the human ear has high spectral resolution at low frequency and low spectral resolution at high frequency. Therefore, in the proposed scheme, a real filter-bank composed of N_f channels equidistantly spaced on the ERB [10] scale was applied to imitate the frequency resolution of human hearing. The bandwidths of the channels in the filter bank are proportional to the center frequencies (see Figure 2). The real filter bank can obtain a good spectral resolution at low frequencies and a good temporal resolution at high frequencies (like the human ear) [15]. Another advantage of the filter bank approach is that each bandpass channel is treated essentially independently, i.e., there are no global spectral constraints on the filter bank outputs [14]. In this specific case, a Hanning window on the frequency side was chosen [2] and the experimental results showed that the type of filter has little influence on the obtained cepstral feature. The output of the j-th channel in the filter bank for the input $\mathbf{s}_{wl\_i}$ is denoted as $\mathbf{s}_{wla\_i}^{(j)}$.

2.5 Hair Cell Transduction

In PLP [13], the cubic-root amplitude compression is combined to approximate the power law of hearing and simulate the nonlinear relation between the intensity of the sound and its perceived loudness [29]. Meddis's hair cell transduction model [17] is incorporated in the MPLPLC model to simulate the rectification, compression, spontaneous firing, saturation effects, and adaptation characteristics of auditory nerve responses [32]. This operation also helps to reduce the spectral amplitude variation of the auditory spectrum, which makes it possible to do the all-pole modeling by a relative low model order [13]. The hair cell transduced version of $\mathbf{s}_{wla\_i}^{(j)}$ is denoted as $\hat{\mathbf{s}}_{wla\_i}^{(j)}$.

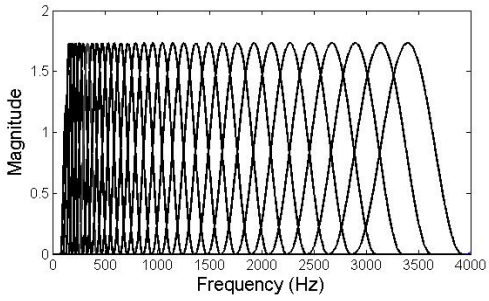


Figure 2. Frequency responses of the filters in the auditory filter bank, with center frequencies equally spaced between 131 Hz and 3400 Hz on the ERB-rate scale.

2.6 Filter Bank Based Energy Calculation

To represent the energy distribution of the music signal on each channel, the energy of the j-th channel for the i-th frame, denoted as $g_i(j)$, is calculated as follows:

[2] http://ltfat.sourceforge.net/doc/filterbank/erbfilters.php

$$g_i(j) = \log \sum_{n=1}^{L_w} \left(\hat{s}_{wla\_i}^{(j)}(n) \right)^2 \quad (2)$$

Here, $\hat{s}_{wla\_i}^{(j)}(n), n = 1, \cdots, L_w$ is the element of the vector $\hat{\mathbf{s}}_{wla\_i}^{(j)}$. Then, the filter bank based energy of the i-th frame is $\mathbf{g}_i = [g_i(1), \cdots, g_i(N_f)]$.

2.7 Autoregressive Modeling

To represent the spectral envelope of the filter bank based energy in a compressed form, the filter bank based energy $\mathbf{g}_i, i = 1, \cdots, N$ are modelled by a pth-order all pole spectrum $\sigma/A_i(z)$, where σ is constant and $A_i(z) = 1+a_{i1}z^{-1}+\cdots+a_{ip}z^{-p}$, using the autocorrelation method [16]. Then, the LPCs of the ith frame are denoted as $\mathbf{a}_i = [a_i(1), \cdots, a_i(p)]$.

2.8 LPC Cepstrum Coefficients Mapping

To reduce the correlation between them [5], the LPCs $\mathbf{a}_i$ are further transformed into (real) LPC cepstrum coefficients, denoted as $\mathbf{c}_i = [c_i(1), \cdots, c_i(p)]$, with the following recursion formula [14]:

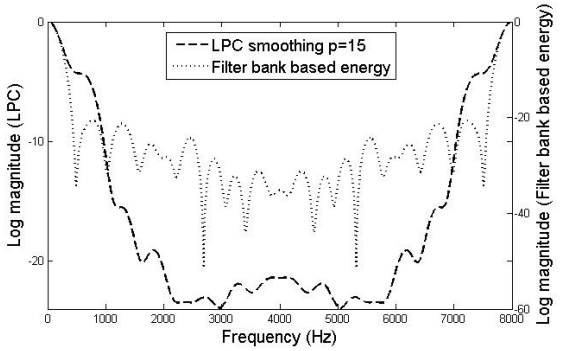
$$c_i(n) = -a_i(n) - \frac{1}{n}\sum_{k=1}^{n-1}(n-k)a_i(k)c_i(n-k) \quad (3)$$

Figure 3(a) and 3(b) show the comparison between the spectrum of filter bank based energy and its LPC smoothing result, and that between the spectrum of filter bank based energy and its cepstrum smoothing result, respectively. It can be seen that first, both the LPC and the corresponding LPC cepstrum can represent the rough change trend of the spectral envelop of the filter bank based energy, and second, the LPC smoothing does not follow the slow variations of the filter bank based energy as well as LPC cepstrum smoothing does. This means that the LPC cepstrum mapping helps to reduce the unnecessary sensitivity that exists in LPC smoothing results.

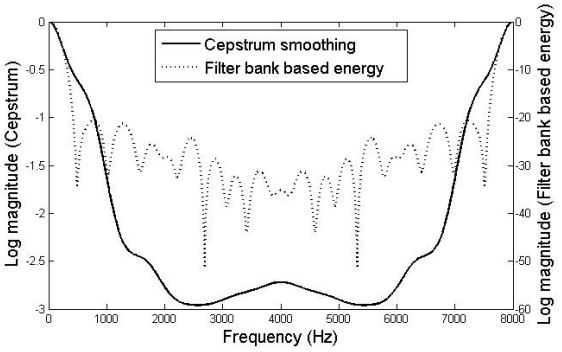
2.9 Cepstrum Liftering

It has been proven that the variability of low quefrency terms is primarily due to variation in transmission, speaker characteristics, and vocal efforts of the human voice [14]. As for the music, the lower quefrency is closely related to the aspect of timbre [19, 21, 30]. So, to boost the degree of timbre invariance of the proposed feature, the liftering window proposed in [14] [see Eq.(4)] is applied to the LPCs first; then, the lower q elements of the result are truncated to get the liftered LPCs denoted as $\hat{\mathbf{c}}_i = \{\hat{c}_i(1), \cdots, \hat{c}_i(p-q)\}$.

$$W_L(n) = \begin{cases} 1 + \frac{p}{2}sin(\frac{\pi n}{p}), & n = 1, 2, \cdots, p \\ 0, & otherwise \end{cases} \quad (4)$$



(a)



(b)

Figure 3. Comparison of spectral smoothing methods.

3. EVALUATION

3.1 Evaluation Preparation

To test the effectiveness of the MPLPLC feature in the pop CSI task, the enhanced Q_{max} method [27] (denoted as $\hat{Q}_{max}$ in this paper) was used to measure the distance between the MPLPLC time series of two pieces of music. The parameters chosen to calculate cross recurrence plots [34] were embedding dimension $m = 15$, time delay (in units) $\tau = 2$ and the maximum percentage of neighbours $\kappa = 0.1$. Furthermore, the parameters used to compute a cumulative matrix Q [26] are the penalty for a disruption onset $\gamma_o = 5$ and the penalty for a disruption extension $\gamma_e = 0.5$.

Two music collections were used. The first one (denoted as Collection_1) comprised 502 pop songs of various styles and genres and 212 cover sets. The average number of covers in each cover set is 2.4, and the distribution of the cover set cardinality has been presented in Figure 4. Western songs and Chinese songs occupy one half of this collection. The second one (denoted as Collection_2) is independent of Collection_1 and comprised 175 songs and 85 cover sets. The cover versions of each cover set in Collection_2 were pop songs performed by the same artist but with different instrumental accompaniments. The materials were obtained from a personal music collection. The identification accuracy and robustness against variation in instrumental accompaniments of the MPLPLC was tested on Collection_1 and Collection_2, in comparison with the

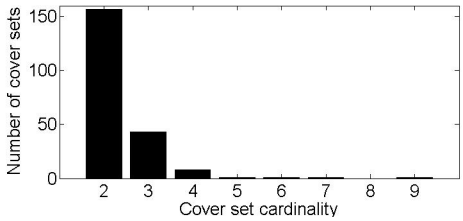


Figure 4. Distribution of the cover set cardinality.

PLP feature [13], CRP feature [19] [3], Ellis's cover song scheme [7] [4], and Serrà's cover song scheme [27] [5]. The parameters of the MPLPLC model have been listed in Table 1, and those of PLP, CRP, Ellis's scheme, and Serrà's scheme are the same as those in [13], [19], [7], and [27], respectively.

Table 1. Parameter setting of MPLPLC feature

Description	Value
Preemphasis parameter μ	0.97
Frame length	464ms
Frame overlap	116ms
Minimum central frequency of auditory filter	133Hz
Maximum central frequency of auditory filter	6856Hz
Number of channels in auditory filter bank N_f	41
LPC order p	16
Number of cepstrum	16
Cepstrum truncate number q	3

3.2 Identification Accuracy

We used each of the 502 songs in Collection_1 as a query and calculated the distance [27] between each query and the remaining 501 songs based on different features. The identification accuracy, in terms of TOP-5, MAP, RANK, and MaRR, obtained from the distance matrices (see Table 2) demonstrated that MPLPLC performed better than the conventional features in the CSI task over Collection_1. One possible explanation for this result is that Collection_1 was composed of pop songs that included a singing voice, and due to the MPLPLC's background in speech recognition, it outperformed the musical facet based features in representing the singing voice. As an example, we studied two versions of the song *Wishing We Last Forever* as performed by Teresa Teng and Faye Wong, respectively. In these two versions, the singing voice is dominant, the instrumental accompaniments are different, and the rhythm is smoothing. The version performed by Teresa Teng includes a national instrument accompaniment, which doesn't conform to the twelve-tone equal temperament. The cross recurrence plots for these two versions based on MPLPLC, CRP [19], beat-synchronous chroma [7] and HPCP [27] have been presented in Figure 5(a)-(d), respectively. We observe that the extended pattern in Figure 5(a), which corresponds to similar sections in two versions, is much more distinct and longer than those in Figure 5(b)-(d). This indicates that first, MPLPLC may outperform the other features in representing the singing voice characteristics, and second, the difference in harmonic information resulting from the difference in instrumental accompaniment affects the performance of PCP-based features.

[3] http://resources.mpi-inf.mpg.de/MIR/chromatoolbox/
[4] http://labrosa.ee.columbia.edu/projects/coversongs/
[5] http://joanserra.weebly.com/publications.html

Table 2. The identification accuracy comparison among MPLPLC and conventional features over Collection_1.

System	Identification accuracy			
	TOP-5	MAP	RANK	MaRR
MPLPLC $+\hat{Q}_{max}$	**738**	**0.9446**	**3.79**	**0.4387**
PLP [13] $+\hat{Q}_{max}$	386	0.4783	58.52	0.2392
CRP [19]$+\hat{Q}_{max}$	525	0.6719	56.48	0.3237
Ellis's [7]	600	0.7489	28.32	0.3507
Serrà's [27]	558	0.7266	28.28	0.3507

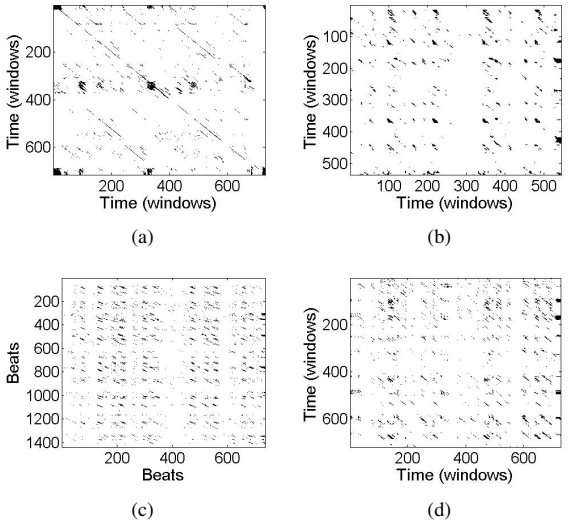


Figure 5. Cross recurrence plot for two versions of *Wishing We Last Forever* as performed by Teresa Teng and Faye Wong based on different features: (a) MPLPLC ($\hat{Q}_{max} = 464.5$), (b) CRP ($\hat{Q}_{max} = 21$), (c) Beat-synchronous chroma ($\hat{Q}_{max} = 61.5$), and (d) HPCP ($\hat{Q}_{max} = 47.5$)

3.3 Robustness against Variation in Instrumental Accompaniments

When compared with classical music, popular music can present a richer range of variation in style and instrumentation [8]. To test the robustness of MPLPLC against variation in style and instrumentation, the identification accuracy in terms of MAP achieved by MPLPLC and by the conventional features were tested and compared with Collection_2. The experimental results shown in Figure 6 indicate that the MPLPLC feature achieves a higher degree of invariance against instrumental accompaniment than the PLP feature [13], CRP feature [19], Ellis's scheme [7], and Serrà's scheme [27]. This phenomenon may also result from the MPLPLC's ability of representing the singing voice.

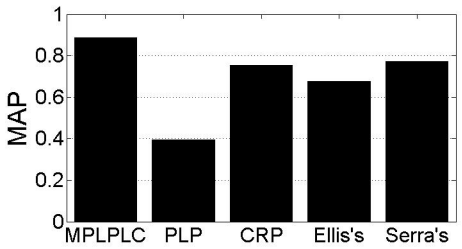


Figure 6. Comparison of robustness against variation in instrumental accompaniments over Collection_2.

3.4 Effect of Cepstrum Mapping and Liftering

To demonstrate the influence of the step LPC cepstrum coefficients mapping and cepstrum liftering on the identification power of of the MPLPLC feature, the identification accuracy based on the MPLP feature, which is obtained by the MPLPLC model without LPC cepstrum coefficients mapping and cepstrum liftering steps; the MPLPC feature, which is generated by the MPLPLC model without cepstrum liftering step; and the MPLPLC feature, have been compared in terms of TOP-5, MAP, RANK, and MaRR over Collection_1 in Figure 7. It can be seen that both LPC cepstrum coefficients mapping and cepstrum liftering help to enhance the identification power of the MPLPC feature.

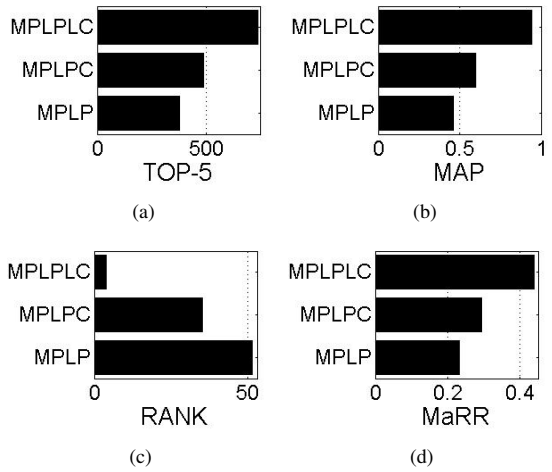


Figure 7. Identification accuracy comparison among MPLP feature, MPLPC feature, and MPLPLC feature, in terms of (a) TOP-5, (b) MAP, (c) RANK, and (d) MaRR over Collection_1.

4. CONCLUSION

We present a new approach, the MPLPLC model, to extract perceptually relevant features from the music signals for pop cover song identification. Here, our main idea is to modify the PLP model, which is a mature technique in the ASR field, by introducing the newest research achievements in psychophysics, such as the time-varying loudness model, auditory filter bank model, and hair cell transduction model, and by taking the difference between speech and music signals into consideration. Furthermore, LPC cepstrum mapping and cepstrum liftering are combined in the proposed model to boost the resulting feature towards timbre invariance. Experimental results over two music collections show that MPLPLC achieves higher identification accuracy and degree of invariance against instrumental accompaniment than the conventional PLP feature and state-of-the-art music theory based features [7, 19, 27] in the CSI task. This means that the mature techniques in ASR may be modified and used in CSI or other MIR fields.

Despite these achievements, there still exists a lot of room for improvement. Since the MPLPLC feature is based on the modification of PLP, which has been successful in the ASR field, it is good at representing singing voice characteristics. As a result, the MPLPLC-based CSI scheme can identify cover versions with a prominent sing voice very well but not those with only instrumental sounds. To solve this problem, in the near future, we will study the SCI scheme, which is based on the fusion of the MPLPLC feature and the musical facet based features (e.g. PCP-based features), which are good at analyzing harmony-based western music. Furthermore, we plan to look into the application of the MPLPLC feature for other MIR tasks, such as structure analysis, cross-domain music matching, and music segmentation.

5. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No. 61271349) and the Natural Science Foundation of Shanghai, China (12ZR1415200).

6. REFERENCES

[1] J. Benesty, M.M. Sondhi, and Y.T. Huang. *Springer Handbook of Speech Processing*. Springer Science & Business Media, 2008.

[2] S. Bleeck, T. Ives, and R.D. Patterson. Aim-mat: the auditory image model in matlab. *Acta Acustica united with Acustica*, 90(4):781–787, 2004.

[3] T.M. Chang, E.T. Chen, C.B. Hsieh, and P.C. Chang. Cover song identification with direct chroma feature extraction from aac files. In *2nd Global Conference on Consumer Electronics*, pages 55–56. IEEE, 2013.

[4] P.J. Clemins and M.T. Johnson. Generalized perceptual linear prediction features for animal vocalization analysis. *The Journal of the Acoustical Society of America*, 120(1):527–534, 2006.

[5] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals*. IEEE New York, NY, USA:, 2000.

[6] J.S. Downie. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[7] D.P.W. Ellis and G.E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *International Conference on Acoustics, Speech and Signal Processing*, pages IV–1429. IEEE, 2007.

[8] D.P.W. Ellis and B.M. Thierry. Large-scale cover song recognition using the 2d fourier transform magnitude. In *The 13th International Society for Music Information Retrieval Conference*, pages 241–246, 2012.

[9] T. Fujishima. Realtime chord recognition of musical sound: A system using common lisp music. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 464–467, 1999.

[10] B.R. Glasberg and B.C.J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138, 1990.

[11] B.R. Glasberg and B.C.J. Moore. A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50(5):331–342, 2002.

[12] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, 2006.

[13] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[14] B.H. Juang, L. Rabiner, and J.G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(7):947–954, 1987.

[15] J.C. Junqua, J.P. Haton, and H. Wakita. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers Boston, USA, 1996.

[16] J. Makhoul. Spectral linear prediction: Properties and applications. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(3):283–296, 1975.

[17] R. Meddis, M.J. Hewitt, and T.M. Shackleton. Implementation details of a computation model of the inner hair-cell auditory-nerve synapse. *The Journal of the Acoustical Society of America*, 87(4):1813–1816, 1990.

[18] B.C.J. Moore. Development and current status of the cambridge loudness models. *Trends in hearing*, 18:1–29, 2014.

[19] M. Müller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662, 2010.

[20] A.V. Oppenheim, R.W. Schafer, J.R. Buck, and other. *Discrete-Time Signal Processing*, volume 2. Prentice-hall Englewood Cliffs, 1989.

[21] F. Pachet and J.J. Aucouturier. Improving timbre similarity: How high is the sky. *Journal of Negative Results in Speech and Audio Sciences*, 1(1):1–13, 2004.

[22] A. Přibilová. Preemphasis influence on harmonic speech model with autoregressive parameterization. *Radioengineering*, 12(3):33–36, 2003.

[23] J. Salamon. *Melody Extraction from Polyphonic Music Signals*. PhD thesis, Universitat Pompeu Fabra, 2013.

[24] J. Serrà. *Identification of Versions of the Same Musical Composition by Processing Audio Descriptions*. PhD thesis, Universitat Pompeu Fabra, 2011.

[25] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, 2008.

[26] J. Serrà, X. Serra, and R.G. ANDRZEJAK. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):111–222, 2010.

[27] J. Serrà, M. Zanin, and R.G. Andrzejak. Cover song retrieval by cross recurrence quantification and unsupervised set detection. *MIREX Extended Abstract*, 2009.

[28] J.O. Smith. *Mathematics of the Discrete Fourier Transform (DFT): with Music and Audio Applications*. Julius Smith, 2007.

[29] S.S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):153, 1957.

[30] H. Terasawa, M. Slaney, and J. Berger. The thirteen colors of timbre. In *Proc. IEEE WASPAA, New Paltz, NY, USA*, pages 323–326, 2005.

[31] T.C. Walters, D.A. Ross, and R.F. Lyon. The intervalgram: An audio feature for large-scale cover-song recognition. In *From Sounds to Music and Emotions*, pages 197–213. Springer, 2013.

[32] D.L. Wang and G.J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[33] C. Xiao. Cover song identification using an enhanced chroma over a binary classifier based similarity measurement framework. In *International Conference on Systems and Informatics*, pages 2170–2176. IEEE, 2012.

[34] J.P. Zbilut, A. Giuliani, and C.L. Webber. Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A*, 246(1):122–128, 1998.