FOREGROUND HARMONIC NOISE REDUCTION FOR ROBUST AUDIO FINGERPRINTING

Matthew C. McCallum

Gracenote Inc.

ABSTRACT

Audio fingerprinting systems are often well designed to cope with a range of broadband noise types however they cope less well when presented with additive noise containing sinusoidal components. This is largely due to the fact that in a short-time signal representation (over periods of $\approx$ 20ms) these noise components are largely indistinguishable from salient components of the desirable signal that is to be fingerprinted. In this paper a front-end sinusoidal noise reduction procedure is introduced that is able to remove the most detrimental of the sinusoidal noise components thereby improving the audio fingerprinting system's performance. This is achievable by grouping short-time sinusoidal components into pitch contours via magnitude, frequency and phase characteristics, and identifying noisy contours as those with characteristics that are outliers in the distribution of all pitch contours in the signal. With this paper's contribution, the recognition rate in an industrial scale fingerprinting system is increased by up to 8.4%.

Index Terms— Audio enhancement, audio fingerprinting

1. INTRODUCTION

Audio fingerprinting addresses the issue of identifying the recording from which a snippet of audio was produced [1]. In this work, specific attention is paid to mobile audio fingerprinting scenarios where an acoustic signal recorded via a microphone is to be identified [2]. This introduces the potential problem of additive acoustic noise to the fingerprinting system. It is also a common use case in audio fingerprinting systems, and as such, there are many robust audio fingerprinting methodologies that are designed accordingly [3–9]. This paper contributes an efficient algorithm to addressing the problems raised by noise sources that are particularly detrimental to the performance of current state of the art audio fingerprinting systems.

A typical fingerprinting system consists of a front end which transforms an audio signal into a compact and unique representation, and a back end in which a matching representation is searched for in a database. While the introduction of additive acoustic noise can prompt careful consideration of in the back end, e.g., [10], much of the literature focuses on contributions that offer robustness to additive acoustic noise at the front end. As a first consideration, audio fingerprinting systems are typically designed to encode features that are largely unchanged by the presence of noise, such as magnitude short-time Fourier transform (STFT) peaks in [11–14], chroma peaks in [15], maximum amplitude Gabor atoms in [16], maximum amplitude wavelets in [4], and differences in magnitude across the frequency and time directions of the STFT [7, 17]. Further research also proposed approaches to further mitigate the effects of noise in these existing fingerprinting systems as in [3, 6, 8, 9, 18].

The aforementioned robust features used for audio fingerprinting systems are typically based on the highest energy components of a given signal representation. Accordingly these features tend to cope with broadband noise sources such as road noise, babble and white noise because they have a high local signal to noise ratio (SNR) with respect to noise energy that is spread across wider areas of a time-frequency signal representation. This in combination with the aforementioned noise mitigation approaches do an excellent job in maintaining fingerprinting performance in a range of broadband noises. These considerations are effective enough that traditional audio noise reduction approaches such as those in [19, 20] don't typically further improve the performance of fingerprinting systems[1]. This lack of improvement is likely the reason behind the almost non-existent amount of literature on traditional noise reduction approaches applied to audio fingerprinting systems.

While fingerprinting systems cope with broadband noise sources well, noise sources that have high energy components concentrated in time and frequency tend to be particularly problematic. These noise types typically include harmonic content such as speech, humming and singing, which are often present in the mobile audio fingerprinting scenarios this work addresses. For example, it is a common scenario where the recording to be fingerprinted is made on a phone in close proximity to vocal activity, with the music to be fingerprinted playing in the background. To address this issue, some literature has applied source separation methods to audio signals prior to audio fingerprinting [9, 18]. These methods rely on computationally intensive processes such as the computation of a similarity matrix [9] or independent component analysis [18].

This paper contributes an efficient front end noise reduction system for music audio fingerprinting systems that mitigates the effects of the aforementioned problematic noise sources without the need for computationally expensive source separation methods. Because these noise sources are statistically very similar to music signal components in a short-time sense, the noise reduction system in this paper employs characteristics of harmonics calculated over the entire duration of sinusoidal signal components. These characteristics can be employed in order to identify components that are not typical of the music signal, and thereafter remove them.

Previous work on isolating sinusoid sources based on their longer time parameters include [21, 22], although the methods employed in this paper are unique in three ways. Firstly, this paper considers the removal of harmonic noise sources for audio fingerprinting systems. Secondly, sinusoids are tracked independently and later grouped into harmonic sets, allowing the removal of both harmonic contour sets and isolated high amplitude sinusoidal components. Thirdly, phase coherence is employed in sinusoid tracking to mitigate the presence of spurious sinusoid contours. The proposed system consists of four important components: Phasor analysis described in Section 2, contour tracing / grouping described in Sections 3 and 4, classification of contours in Section 5 and removal of contours in Section 6. The output of each of these stages is displayed in Fig. 1.

[1] In fact some of the fingerprint noise mitigation approaches resemble traditional noise reduction methods, e.g., the binary mask in [8].

2. PHASOR ANALYSIS

Here a signal representation is considered that will allow signal (i.e., music) and harmonic noise characteristics to be calculated in order to identify unwanted noise in Section 5, and remove the unwanted noise in Section 6. In this work the STFT is defined as,

$$X\left[k,m\right]=\sum_{n=mM}^{mM+N-1}x\left[n\right]w\left[n-mM\right]e^{\frac{-i2\pi nk}{K}}, \quad (1)$$

where $x[n]$ represents a time domain audio signal and $w[n]$ the windowing function. M refers to the increment in samples between windows, N the windowing length and K the number of frequency bins in the discrete Fourier transform.

While pitch contours could be approximated based directly on amplitude peaks in the STFT in some scenarios, in order to effectively remove a harmonic component its frequency, phase and magnitude must be known with greater accuracy than the STFT offers. Here, the instantaneous frequency, phase and amplitude of signal phasors are considered, calculated from the STFT as,

$$\omega_{k,m}=\frac{2\pi k}{K}+\frac{\left(\angle X\left[k,m\right]-\angle X\left[k,m-1\right]-\frac{2\pi Mk}{K}\right)\bmod 2\pi}{M}, \quad (2)$$

$$\phi_{k,m}=\angle X\left[k,m\right]+\angle W\left(\omega_{k,m}\right), \quad (3)$$

$$A_{k,m}=\frac{2\left|X\left[k,m\right]\right|}{\left|W\left(\omega_{k,m}\right)\right|}, \quad (4)$$

respectively. Here, $W(\omega)$, refers to the discrete-time Fourier transform of the windowing function $w[n]$. The set of frequency, phase and amplitude parameters describing each phasor will be denoted as $\rho_{k,m}$.

This provides a phasor for every frequency bin in the STFT, however, many represent the same harmonic component. In this work only phasors fitting the following conditions are considered,

$$\frac{A_{k,m}}{\max\limits_{k,m}\left|X[k,m]\right|}>\alpha, \quad (5)$$

$$\left|\frac{\left(\angle X\left[k,m\right]-\angle X\left[k,m-1\right]-\frac{2\pi Mk}{K}\right)\bmod 2\pi}{M}\right|<\frac{\pi}{K}. \quad (6)$$

The total set of instantaneous phasors that fit these conditions will be denoted, P. During the process of grouping P into harmonic contours it will also be useful to denote the set of phasors currently grouped into pitch contours as P^{-}, and those remaining as P^{+}.

In order to group the phasors in P into pitch contours that represent the fundamental and harmonics of a music or noise source, a pitch contour tracking procedure reminiscent of [23] is employed, with some differences. As opposed to tracing peaks in a salience function, here the fundamental and harmonics of a pitch contour are traced individually based directly on the instantaneous peaks in P, this allows not only amplitude and frequency continuity along contours, but also phase continuity. In this respect, there are two levels of grouping hierarchy, the grouping of phasors $\rho_{k,m}$ into pitch contours (described in Section 3), and the grouping of pitch contours into harmonic sets (described in Section 4). As such, the notation $\mathrm{C}^{l,h}$ is used to denote an individual pitch contour, where l denotes the harmonic number ($l=0$ corresponding to the fundamental) and h indexing the harmonic set.

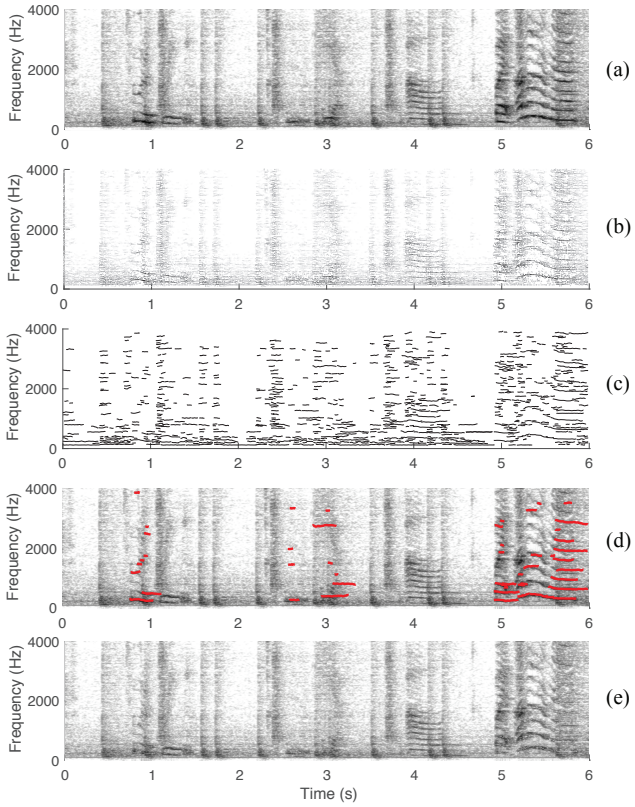

Fig. 1. Output of each of the stages of the noise reduction system. (a) Spectrogram of a snippet of "Brave" by Sara Bareilles corrupted by the colloquial speech utterance "Two to three weeks... yes... umm, but other than that...". (b) Frequency time locations of all phasors satisfying (5,6), amplitude of each phasor is indicated by the darkness of each dot. (c) All contours traced via methods in Section 2. (d) Outlier contours and their harmonics isolated via methods in Section 5, overlaid on the original spectrogram. (e) Enhanced spectrogram after complex spectral subtraction of outlier contours.

3. CONTOUR TRACING

The tracing of any contour, $\mathrm{C}^{l,h}$, begins with a root phasor, $\rho_{r,m}^{root}$. From this root phasor, a pitch contour is traced forwards and then backwards in time to find a set of phasors in P^{+} that adhere to the following amplitude, frequency and phase constraints. Given a phasor that has been selected in the previous frame, $\rho_{s,m-1}$, a phasor in the following frame, $\rho_{k,m}$, is added to $\mathrm{C}^{l,h}$ given,

$$\left|\omega_{s,m-\mu}-\omega_{k,m}\right|<\Delta_f, \quad (7)$$

$$\frac{A_{k,m}}{A_{r,m}}>\Delta_A, \quad (8)$$

$$\min_{\psi}\left|\left(\psi 2\pi-\phi_{k,m}+\phi_{s,m-\mu}+\omega_{s,m-\mu}M\right)\bmod 2\pi\right|<\Delta_\phi. \quad (9)$$

Each of these constraints describes limits on frequency continuity, amplitude decay and phase continuity, respectively. Here $\psi\in\{0,1\}$, and μ refers to the step size of contour tracing - positive for

tracing forwards, and negative for tracing backwards (here we only consider $\mu \in \{-1, 1\}$). The set of all phasors in P^{+} satisfying the aforementioned constraints is described as $\mathrm{P}^{l,h,m}$ when tracing a harmonic. Of all phasors in $\mathrm{P}^{l,h,m}$ the phasor that is selected and added to $\mathrm{C}^{l,h}$ is denoted $\rho_m^{l,h}$, and its parameters denoted, $\omega_m^{l,h}$, $\phi_m^{l,h}$ and $A_m^{l,h}$. Note that frequency bin k is dropped for notational simplicity.

The phasor, $\rho_m^{l,h}$, is selected as that in $\mathrm{P}^{l,h,m}$ with the minimum complex distance from the phase advanced previous frame, i.e., that at the index k which solves,

$$\underset{k \in P^{l,h,m}}{\arg\min} \left| A_{k,m} e^{i\phi_{k,m}} - A_{s,m-\mu} e^{i\left(\omega_{s,m-\mu} M + \phi_{s,m-\mu}\right)} \right|. \quad (10)$$

Once selected, the selected phasor is removed from P^{+} and added to P^{-}. If no suitable phasors are found, a hypothetical phasor is synthesized as,

$$\rho_{s,m} = \{\omega_{s,m-\mu}, \phi_{s,m-\mu} + |M\mu|\, \omega_{s,m-\mu}, A_{s,m-\mu}\} \quad (11)$$

which then continues the contour until a stopping criterion is reached. That is, the tracing procedure (forwards or backwards) stops when no suitable phasors are found for O_{skip} frames. Once traced forwards and backwards from the root phasor $\rho_{r,m}^{root}$, the set $\mathrm{C}^{l,h}$ consists of a set of phasors selected from P in addition to 0 or more synthesized phasors via (11), over a continuous set of frame indices, m, starting at frame, $\acute{m}^{l,h}$, and ending at frame, $\grave{m}^{l,h}$.

4. HARMONIC GROUPING

Using the contour tracing procedure described in Section 3, first a fundamental contour, $\mathrm{C}^{0,h}$, is traced. In this case, the root phasor, $\rho_{r,m}$, is selected as that with the highest magnitude in P^{+}. Once $\mathrm{C}^{0,h}$ has been obtained, additional contours are searched for at each of its harmonics, $l \in [1, L]$. The root phasor for a given harmonic contour is selected as the maximum amplitude phasor $\rho_{k,m}$ that has a frequency component fitting the condition,

$$l\omega_m^{0,h} - \Omega_{dev} \leq \omega_{k,m} \leq l\omega_m^{0,h} + \Omega_{dev} \quad (12)$$

at any m, for which $\mathrm{C}^{0,h}$ exists. Note that when tracing harmonics, the condition in (12) is added as an additional condition to the set of conditions in (7,8,9) for forming the set $P^{l,h,m}$ at frame m.

The set of all contours and its harmonics form a harmonic contour set H^{h}. Using the aforementioned procedures, harmonic contour sets are traced until $\nu\%$ of all phasors in P are also included in any H^{h}.

5. CONTOUR CLASSIFICATION

In this section a method is proposed to identify noise contours which may be then removed from the signal using the method in Section 6. Many common harmonic noise sources such as speech, whistling, sirens, etc. are monophonic and as such typically contain many less pitch contours / continuous sinusoidal components than a typical music signal. The approach proposed here intends to identify pitch contours that are not typical of the signal observed (i.e., statistical outliers) and thereafter remove them from the signal.

Based on the information contained in each contour, i.e., $\omega_m^{l,h}$, $\phi_m^{l,h}$, and $A_m^{l,h}$, a number of contour characteristics may be calculated. Here the following features are considered for each contour,

$$\hat{A}^{l,h} = \max_m A_m^{l,h} \quad (13)$$

$$\hat{\omega}^{l,h} = \frac{1}{\grave{m}^{l,h} - \acute{m}^{l,h} - 1} \sum_{m=\acute{m}^{l,h}+1}^{\grave{m}^{l,h}} \left| \omega_m^{l,h} - \omega_{m-1}^{l,h} \right| \quad (14)$$

$$\hat{\xi}^{l,h} = \frac{\sum_{m=\acute{m}^{l,h}+1}^{\grave{m}^{l,h}} \left(A_m^{l,h}\right)^2}{\sum_{m=\acute{m}^{l,h}+1}^{\grave{m}^{l,h}} \left(\delta_m^{l,h}\right)^2} \quad (15)$$

Where $\delta_m^{l,h}$ is the "tracing noise" for each phasor in $\mathrm{C}^{l,h}$ and is calculated similarly to (10)

$$\delta_m^{l,h} = \left| A_m^{l,h} e^{i\phi_m^{l,h}} - A_{m-1}^{l,h} e^{i\left(\omega_{m-1}^{l,h} M + \phi_{m-1}^{l,h}\right)} \right|. \quad (16)$$

These parameters can be used to then identify pitch contours with unusual characteristics as noise. In practice, depending on the application and signal, a variety of other functions of $\omega_m^{l,h}$, $\phi_m^{l,h}$, and $A_m^{l,h}$ may also be useful, for example, in the case where the desired signal is western music, values of median $\omega_m^{l,h}$ may be useful in identifying outliers that do not adhere to the key of the audio snippet.

In this work (13) was found to be the most important attribute in identifying noise that is detrimental to audio fingerprinting systems for two reasons. Firstly, the desired signal - music, over any small frequency range, usually contains a relatively homogeneous distribution of $\hat{A}^{l,h}$, with low variance, especially in the case of more modern music recordings that undergo heavy use of multi-band compression, and so outliers in $\hat{A}^{l,h}$ are not typical of a music signal. Secondly, considering the fingerprinting schemes of [11–14], it is pitch contours with points of considerably higher $\hat{A}^{l,h}$ that are most problematic to the fingerprinting algorithm, therefore by identifying these as noise and removing them, the most significant improvements in audio fingerprinting performance are attained.

In order to identify noise pitch contours, first a threshold is employed, $\hat{\xi}^{l,h} > \Delta_{\hat{\xi}^{l,h}}$, over all contours, to prevent spurious pitch contours being identified as noise. Next outliers in $\hat{A}^{l,h}$ and $\hat{\omega}^{l,h}$ are identified by observing z-scored contours that exceed empirically derived thresholds, i.e., $\Delta_{\bar{A}}$ and $\Delta_{\bar{\omega}}$, respectively. Here a bar is used to denote z-scored variables calculated as,

$$\bar{y} = \frac{y - \eta}{\sigma_y}, \quad (17)$$

where η is the median calculated over all values of y and σ the median absolute deviation [24]. These statistics are employed to provide robustness to the expected outliers in the music/noise signal. Specifically outliers are expected at higher values of $\bar{A}^{l,h}$ and $\hat{\omega}^{l,h}$. This idea is motivated by the fact that the tracing procedure in Section 2 operates by grouping only the $\nu\%$ of phasors that form consistent contours each starting with the highest remaining amplitude phasor $\rho_{k,m}$ in the set of ungrouped phasors. Similarly for $\hat{\omega}^{l,h}$, it is expected that due to the nature of music signal, there will be a large number of contours with a very small $\hat{\omega}^{l,h}$, that result from sustained musical notes. As such, the distribution of $\hat{\omega}^{l,h}$ will be concentrated around its lower bound of 0, with outliers at higher values.

It may be noted that the scale parameters $\sigma_{\bar{A}}$ and $\sigma_{\bar{\omega}}$ are biased estimators for scale and dependent on the distribution of the feature for which it is calculated. However, as the thresholds $\Delta_{\bar{A}}$ and $\Delta_{\bar{\omega}}$ are empirically derived, the effect of any bias in $\sigma_{\bar{A}}$ and $\sigma_{\bar{\omega}}$ may be considered absorbed into these thresholds.

Each of the z-scored values calculated via (17) that exceed the aforementioned thresholds and contain a minimum of τ phasors are

added to the noise contour set, N. In addition, any contours in the same harmonic set H^h of a contour $\mathrm{C}^{l,h}$ that is in the noise set N, are also added to N if they too exceed the minimum contour length .

Outlier detection based on the observed statistics of $\hat{A}^{l,h}$ and $\hat{\omega}^{l,h}$ is a relatively simple calculation that requires a modest amount of computational resources. This makes the aforementioned approach suitable for the mobile devices on which audio is often recorded for fingerprinting applications.

6. CONTOUR REMOVAL

Once identified, given each contour is considered an accurate representation of a deterministic signal, the optimal removal of contours is performed by complex spectral subtraction [21]. Here the approach is to simply synthesize a noise spectrum by summing,

$$d[n] = 2A_{k,m}\cos\left(\omega_{k,m}n + \phi_{k,m}\right), \tag{18}$$

over all positive frequency sets $\{A_{k,m}, \omega_{k,m}, \phi_{k,m}\}$ in N. The spectrum $D[k,m]$ is then computed via (1) and subtracted from $X[k,m]$, providing the noise reduced spectrum $Y[k,m]$. $Y[k,m]$ can be re-synthesized in the time domain, or used directly in calculating features for audio fingerprinting. It is important to note that the interpolation in frequency and phase performed in (2,3,4) is crucial to the accurate removal of deterministic signal components.

7. RESULTS

Here the performance of the proposed noise reduction front end is evaluated both in terms of objective signal quality and in terms of improvement in fingerprinting performance. For both cases a dataset of 500 6 second snippets of audio from 167 popular songs is used to create fingerprint queries. Each recording was combined with each of pink noise and talking noise and thereafter convolved with an impulse response that was measured on an iPhone 6 Plus. Noise was added at global SNRs ranging from -10 dB to 10 dB at 2dB increments. The talking noise used was taken from the well known TIMIT database.

In all experiments, the work described in this paper was configured with the following parameters: a sampling rate of 8 kHz, $w[n]$ is a Hamming function, Δ_f was set to the bandwidth of $W(\omega)$, $N = 160$, $M = 16$, $K = 640$, $\Omega_{dev} = 0.062$, $\alpha = 2.13 \times 10^{-6} W(0)$, $O_{skip} = 10$, $L = 13$, $\tau = 50$, $\Delta_A = 0.3$, $\Delta_p hi = 1.0$, $\Delta_{\hat{\xi}^{l,h}} = 1.0$, $\Delta_{\bar{A}} = 6.68$, $\Delta_{\bar{\omega}} = 8.0$, $\nu = 30\%$.

As an objective quality measure, the normalized covariance metric (NCM) is employed [25]. This metric is configured to measure 64 linearly spaced bands from 0 Hz to 4 kHz all with equal weightings, as the fingerprinting system employed in this paper is largely impartial to frequency locations of peaks.

It can be seen in Fig. 2 that at SNRs of 0 dB and below there is a clear improvement the NCM value indicating that there is an objective improvement in signal quality. The reduction in speech noise in the observed examples are usually localised to particularly loud speech phonemes, typically at frequencies < 2 kHz. These local improvements are important for fingerprinting algorithms that only need reliable content over a subset of time and frequency, but are de-emphasized in a metric such as the NCM here which equally weights the effect of distortions at all times and frequencies.

To demonstrate the utility of the algorithm proposed in this paper, it is employed as a preprocessor to the industrial scale audio fingerprinting system in [14]. For the experiment here, the fingerprint index consisted of 120k songs. All 500 6 second song snippets

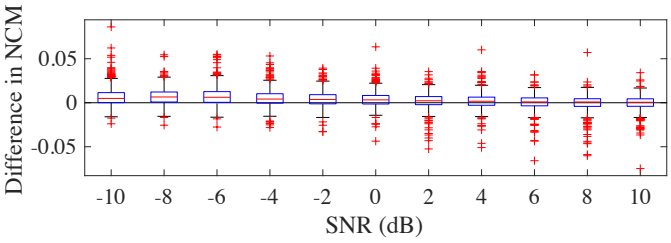

Fig. 2. The difference in NCM realized by employing the methods described in this paper. Differences are computed for each individual track corrupted by talking noise at the range of SNRs shown.

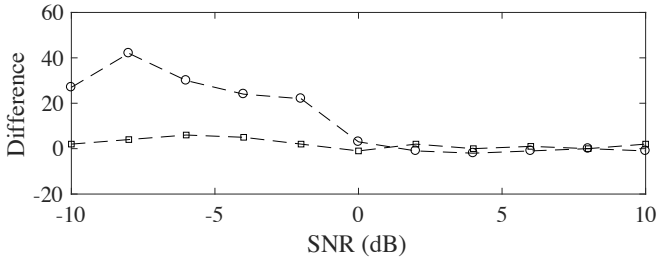

Fig. 3. The difference in number of successfully identified 6 second fingerprint queries realized by employing the methods described in this paper for samples corrupted by ∘ speech, and □ pink noise.

combined with each of the 2 noise types at all 11 SNRs were run as queries against this fingerprinting system. The difference in number of correctly identified queries for each scenario with and without the noise reduction methods described in this paper are shown in Fig. 3. It can be seen that for SNRs below 0 dB there is a significant improvement in the number of queries recognized when corrupted by talking noise. As expected there is little to no improvement for the pink noise case due to its lack of harmonic content. However, it is promising to see that there is no degradation to the fingerprinting performance as might be suspected due to the unwanted removal of desirable harmonic content. This is likely due to the relatively homogeneous nature of pitch contour amplitudes in music signals.

8. CONCLUSION

This work introduced a noise reduction system that mitigates the adverse affects of harmonic noise sources on fingerprinting systems. It was shown that by tracking a number of highest amplitude pitch contours over an audio signal, a distribution of pitch contour parameters may be used to identify and thereafter remove some of the most abnormal pitch contours for that signal. By removing such outliers in maximum contour amplitude and mean frequency change, an improvement in both signal quality (indicated via the NCM) and in fingerprinting recognition rate can be obtained. In future work these methods could be improved upon via further fine tuning of the parameters, improved sinusoid parameter tracing methods (e.g., such as those in [26]), investigation into alternative pitch contour parameters, or more sophisticated pitch contour classification methods such as support vector machines or deep learning methods.

9. REFERENCES

[1] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma, "A review of audio fingerprinting," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 41, no. 3, pp. 271–284, 2005.

[2] Vijay Chandrasekhar, Matt Sharifi, and David A Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications.," in *ISMIR*, 2011, pp. 801–806.

[3] Mansoo Park, Hoi-Rin Kim, and Seung Hyun Yang, "Frequency-temporal filtering for a robust audio fingerprinting scheme in real-noise environments," *ETRI journal*, vol. 28, no. 4, pp. 509–512, 2006.

[4] Shumeet Baluja and Michele Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern recognition*, vol. 41, no. 11, pp. 3467–3480, 2008.

[5] Chih-Yi Chiu, Dimitrios Bountouridis, Ju-Chiang Wang, and Hsin-Min Wang, "Background music identification through content filtering and min-hash matching," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2414–2417.

[6] Wooram Son, Hyun-Tae Cho, Kyoungro Yoon, and Seok-Pil Lee, "Sub-fingerprint masking for a robust audio fingerprinting system in a real-noise environment for portable consumer devices," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 1, 2010.

[7] Xavier Anguera, Antonio Garzon, and Tomasz Adamek, "Mask: Robust local features for audio fingerprinting," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*. IEEE, 2012, pp. 455–460.

[8] Bob Coover and Jinyu Han, "A power mask based audio fingerprint," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1394–1398.

[9] Hyoung-Gook Kim, Hye-Seung Cho, and Jin Young Kim, "Robust audio fingerprinting using peak-pair-based hash of non-repeating foreground audio in a real environment," *Cluster Computing*, vol. 19, no. 1, pp. 315–323, 2016.

[10] Kimberly Moravec and Ingemar J Cox, "A comparison of extended fingerprint hashing and locality sensitive hashing for binary audio fingerprints," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011, p. 31.

[11] Avery Wang, "An industrial strength audio search algorithm.," in *ISMIR*, 2003, pp. 7–13.

[12] Reinhard Sonnleitner and Gerhard Widmer, "Robust quad-based audio fingerprinting," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 409–421, 2016.

[13] Michaël Betser, Patrice Collen, and Jean-Bernard Rault, "Audio identification using sinusoidal modeling and application to jingle detection.," in *ISMIR*, 2007, pp. 139–142.

[14] Jinyu Han and Bob Coover, "Audio fingerprinting," Mar. 15 2016, US Patent 9,286,902.

[15] Mani Malekesmaeili and Rabab K Ward, "A local fingerprinting approach for audio copy detection," *Signal Processing*, vol. 98, pp. 308–321, 2014.

[16] Courtenay V Cotton and Daniel PW Ellis, "Audio fingerprinting to identify multiple videos of an event," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2386–2389.

[17] Jaap Haitsma and Ton Kalker, "A highly robust audio fingerprinting system.," in *ISMIR*, 2002, pp. 107–115.

[18] Wei Han, Songbin Zhou, Chang Li, Yisen Liu, and Zhe Liu, "Blind source separation for a robust audio recognition scheme in multiple sound-sources environment," *Spectrum*, vol. 1, pp. 1–5, 2015.

[19] Robert McAulay and Marilyn Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[20] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[21] Matthew McCallum and Bernard Guillemin, "Accounting for deterministic noise components in a MMSE STSA speech enhancement framework," in *Communications and Information Technologies (ISCIT), 2012 International Symposium on*. IEEE, 2012, pp. 169–174.

[22] Rachel M Bittner, Justin Salamon, Slim Essid, and Juan Pablo Bello, "Melody extraction by contour classification.," in *ISMIR*, 2015, pp. 500–506.

[23] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[24] Peter J Rousseeuw and Christophe Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[25] Inga Holube and Birger Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[26] Matthew McCallum and Bernard J Guillemin, "Joint stochastic-deterministic Wiener filtering with recursive Bayesian estimation of deterministic speech," in *Interspeech*, 2013, pp. 460–464.