

MUSIC AND NOISE FINGERPRINTING AND REFERENCE CANCELLATION APPLIED TO FORENSIC AUDIO ENHANCEMENT

ANIL ALEXANDER¹, OSCAR FORTH¹ AND DONALD TUNSTALL²

¹ Oxford Wave Research Ltd, United Kingdom
{[anil](mailto:anil@oxfordwaveresearch.com)|[oscar](mailto:oscar@oxfordwaveresearch.com)}@oxfordwaveresearch.com

² Digital Audio Corporation, USA
dtunSTALL@dacaudio.com

In this paper, we present a novel application, of audio fingerprinting and reference cancellation, to forensic audio enhancement. In audio surveillance recordings, it is common to find the speech of interest masked by the speech of other non-target speakers in the room or obscured by interfering music or television noise, as well as other noises such as banging, clanging or slamming, in the acoustic environment. These noises can drown the speech of interest and make it difficult to understand easily or be accurately transcribed. When two recordings of a particular acoustic event are available, two-channel adaptive filtering or reference cancellation is a highly effective tool for enhancement. However, the two recordings have to be painstakingly manually aligned, and this is difficult to do by eye or ear. We present an approach using landmark-based acoustic fingerprinting to identify, automatically align, and subtract reference sounds and bring the speech of interest to the forefront.

1 INTRODUCTION

In this study, we explore the extent to which it is possible to reduce or remove, from contemporaneous recordings made in the same acoustic environment, interfering music, background noises, and speech of other speakers, to bring the voice of the main speaker to the forefront. We have previously presented the application of acoustic fingerprinting and music signal cancellation for the forensic enhancement of audio containing speech masked by background music [Alexander and Forth, 2010]. In this paper, we extend this approach to align and subtract recordings made from two independent audio recorders using landmark-based audio fingerprinting. This has potentially significant applications in audio deployments and subsequent enhancement of surveillance audio recordings.

2 BACKGROUND

The ability to ‘fingerprint’ a section of audio, to identify the source material present in it and to accurately time-align and ‘subtract’ the source material will allow for significant improvement in the intelligibility of the target speech present in the audio.

Reference signal cancellation applied to music and television noise is a challenging problem, which requires the following to be precisely identified: the exact moment in time, within the file being analysed,

that the song or music begins, and the original song or music file playing in the background. Once these are identified, a noise- and distortion-robust signal cancellation algorithm can be applied to remove or reduce the music while mostly leaving the target speech intact. Applying reference cancellation is usually a painstaking manual task, and the recordings have to be precisely aligned by eye and ear.

Music identification has been an area of recent commercial interest with the proliferation of personal handheld mobile devices. Shazam (Wang 2003), MusicDNS, AudioID, etc., are examples of popular music identification systems. The basic idea behind music identification is that a user, often in noisy, distorted or otherwise poor recording conditions, can record a short segment of music playing in his environment and send through to a recognition server for comparison against a database of songs. The recognition server then extracts discriminative features from this short clip and compares it with a pre-indexed database of songs and lists the most probable candidates for the song. The other applications of such acoustic fingerprinting are in identifying tunes, songs, videos, advertisements, radio broadcasts, etc.

Audio fingerprinting of background noises (not necessarily only music or television), followed by subtraction of the reference channel, can be used in order to reduce the interference caused by the background sounds.

3 PROPOSED APPROACH

We propose a two stage approach to reducing the effect of interfering music, television or noise using firstly landmark-based acoustic fingerprinting to identify, and automatically align reference sounds and a signal reference cancellation algorithm technique to subtract the reference sounds to bring the speech of interest to the forefront. The landmark-based audio fingerprinting is explained in detail in Section 3.1.

The typical approach to reference signal cancellation, uses the normalized least mean squares (LMS) algorithm, commonly also used in echo cancellation (Benesty et al, 1997). In order to cancel the interfering sound, the ‘reference’ recording, ideally containing only the correct, pristine interfering audio source, with no room acoustics, needs to be time-aligned, within a few milliseconds, to the position in the recording being enhanced, known as the ‘primary’ recording. The timing information for achieving the required alignment is provided by the landmark-based audio fingerprinting algorithm.

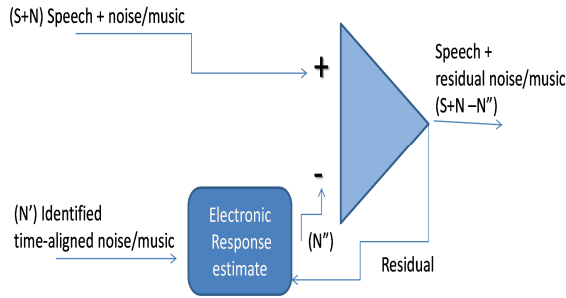


Figure 1: Reference cancellation using identified and time-aligned reference audio

Though conceptually simple, the robust cancellation of one signal from another is a problem requiring complex mathematical modelling to achieve, especially in rooms with high reverberation time. Unfortunately, simple subtraction of the aligned reference from the primary audio is rarely sufficient, on its own, to achieve effective cancellation, even with perfect amplitude and phase matching, due to the fact that the acoustic reverberations of an interfering sound are often nearly as loud as the direct-path sound itself.

The LMS algorithm, coupled with a sufficient length finite impulse response (FIR) filter, to accurately model both the room acoustics and the combined frequency response of the playback and recording equipment, creates a filter that gets applied to the reference audio to properly match it to the primary audio prior to

subtraction (Figure 1). The reference signal cancellation algorithm is explained in Section 3.2.

The resultant processed audio contains mostly the intact speech of the target speaker in the foreground, and the attenuated music/noise in the background, as illustrated in Figure 2.

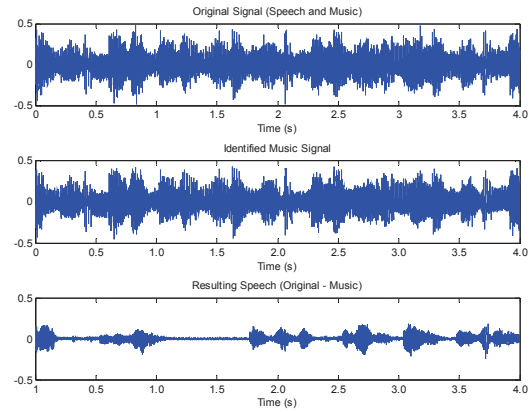


Figure 2: Waveforms of the original signal, subtracted identified music signal and the result of music signal cancellation

3.1 Landmark-based audio fingerprinting

We use a landmark-based acoustic fingerprinting algorithm (Ellis 2009) in order to obtain the start position of the reference track. Wang (2003) lists the following attributes for an acoustic fingerprint— the hashes need to be temporally localized, translation invariant, robust and sufficiently entropic. Spectral peaks, which are observable in music, allow for precise temporal localization and are robust to noise and transmission distortions.

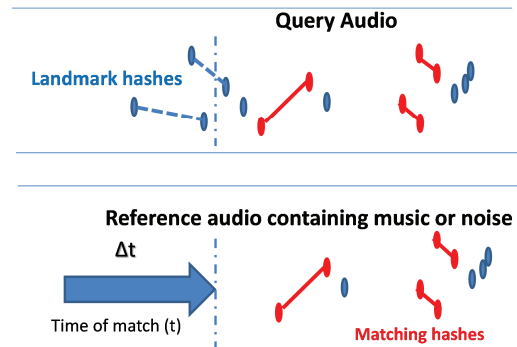


Figure 3: Estimated time alignment by matching landmark constellations

The following steps are involved in landmark-based audio fingerprinting (illustrated in Figure 3).

- Some peaks in a spectrogram are chosen based on having higher energy than their neighbouring ones.
- The spectrogram is then reduced into a 'constellation map' containing spectral peaks.
- From this map, pairs of peaks are selected as landmark 'hashes' that provide reference anchor points in time and frequency.
- Landmark hash extraction is performed on the query audio.
- The constellation maps are then compared to obtain the position in time when some of the hashes match, between the query and reference audio.
- The file with the largest number of hash matches is selected as the reference audio file.
- An accurate estimate of the time of match is also returned by this algorithm.

3.2 Reference Signal Cancellation

As illustrated in Figure 1, internal to the process of reference signal cancellation is the *electronic response estimate*, a simulation of the acoustic environment for the original sound waves, which is typically implemented using a Finite Impulse Response (or 'FIR') filter operating in the time domain, usually at a sample rate of 16000 Hz. This provides approximately 7 kHz of usable audio bandwidth and works very well for speech applications.

In a time-domain implementation, the FIR filter coefficients are typically calculated using either the least-mean-square (LMS) algorithm or its alternative normalized implementation (NLMS). With either LMS or NLMS, the complete set of FIR coefficients, $h(i)$, initially all set to zero, is updated at each sample interval as follows:

$$h_n(i+1) = h_n(i) + \Delta h_n(i), \quad (1)$$

where $h_n(i)$ is the n^{th} coefficient for time sample i .

The update increment, $\Delta h_n(i)$, is computed by the LMS algorithm in the following manner:

$$\Delta h_n(i) = \mu \cdot e(i) \cdot x(i-n) \quad (\text{LMS}) \quad (2)$$

Where $e(i)$ is the Residual audio, or $(S + N - N')$, and $x(i-n)$ is the sample of Reference audio in the FIR data pipeline that is associated with coefficient n . The parameter μ specifies the adaptation convergence rate;

as μ is made larger, the filter will adjust itself in larger steps each sample interval, *i.e.*, converge more rapidly.

The NLMS algorithm, when used, calculates the update increment slightly differently as follows:

$$\Delta h_n(i) = \mu' \cdot e(i) \cdot x(i-n) \quad (\text{NLMS}) \quad (3)$$

Here, the specified μ is scaled to the value μ' with inverse proportion to the average input signal power, or:

$$\mu' = \frac{N \cdot \mu}{\sum_{k=0}^{N-1} x^2(i-k)} \quad (4)$$

where N is the size of the FIR filter, generally expressed in units of 'taps', and the denominator is the sum of the squares of all data samples currently in the filter.

As a rule, the reference signal canceller must have a sufficiently large number of taps to simulate all of the significant acoustic paths traversed by the interfering sound waves (such as those from a television set), not only the direct path waves but also the *reverberations*.

Consider the simplified example in Figure 4:

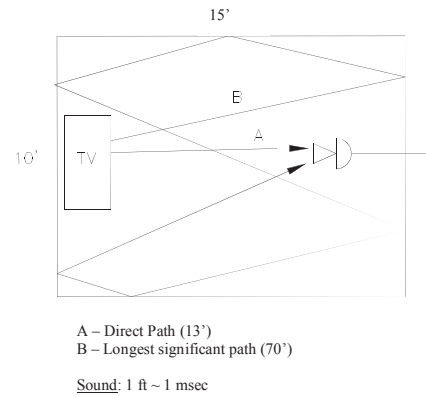


Figure 4: Illustration of Room Acoustics

The largest *significant* reverberation path in this figure is B, which is 70 feet, or 70 milliseconds long (due to sound traveling approximately 1 foot per millisecond in air). Therefore, in order for the FIR filter to accurately simulate this acoustic environment, it must have at least 70 milliseconds of memory.

At a sample rate of 16000 samples per second, each delay in the FIR filter corresponds to $1/16000$ s or 62.5 μ s. The number of delays required, and hence the *minimum* filter size, for this example is $70\text{ms}/0.0625\text{ms} = 1120$ taps.

A simple way to estimate the minimum filter size, to cancel the most significant reverberations in a given room, is to take the largest dimension of the room (in feet) and multiply it by five. For example, a 20' by 15' room would require $(20 \text{ ft} \times 1 \text{ ms/ft} \times 5)$ 100 ms to achieve good reduction of an interfering sound source. At a sampling rate of 16 kHz, this would correspond to a 1600-tap filter. Typically, even larger filters (e.g. 2048, 3072, or 4096 taps) are used, in order to provide additional cancellation of the longer reverberation paths, thereby improving the final results.

4 MUSIC AND NOISE REFERENCE CANCELLATION

4.1 Music Cancellation

For the music cancellation experiments, we used a small reference database of music containing 50 songs. This database consists of popular genres of music, namely pop, rock and instrumental music. The reference music files used for these experiments had a sampling rate of 44100 Hz, a bit-rate of 16 bits and were uncompressed Microsoft wave files. The test recordings were made in the following acoustic environments, with samples from the reference database playing: in a moving (diesel) vehicle, with music playing and windows closed and open, in an office room with music playing and in a garage.

In Figure 5 we can observe the effect of the reference cancellation on the original signal that contains both speech and music. The landmarking algorithm correctly identified the song playing as well as a reasonably precise estimate of the moment in time that it was present in the original signal. This identified music signal was then passed to the reference cancellation algorithm that significantly reduced the background noise.

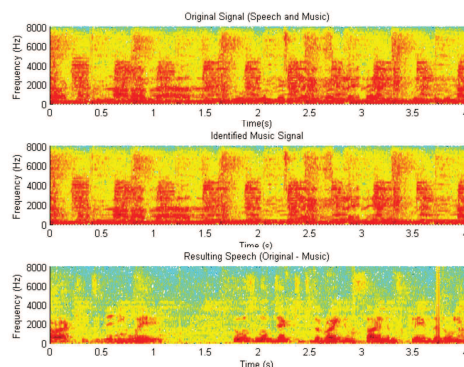


Figure 5: Spectrograms of the original signal, identified music signal, and the result of music signal cancellation. This process was repeated with all of the test recordings and it was observed that:

- The test recordings containing music yielded a high number of hash pairs that were robust to varying interferences such as background noise, room acoustics and engine noise.
- Based on the number of hash matches, it was possible to correctly identify the song or that was playing.
- The time alignment was generally accurate to within the analysis window sizes used in the landmarking algorithm and this variation was generally within the filter taps scope of the reference cancellation algorithm.

In the example illustrated in Figure 6 the test recording that contained Alanis Morissette's song "Uninvited" was compared against two different versions of the song, one a 'live' version and the other a studio version almost identical in instruments, beat, and timing, by the same artist. The landmarking algorithm was correctly able to identify the studio version as the correct one. Note that the dashed red lines represent the landmark hashes in the query and reference audio and the solid white lines represent the hashes that were matched in both audio files.

We thus observe that the fingerprinting algorithm is highly specific to the exact source music that was playing in the acoustic environment.

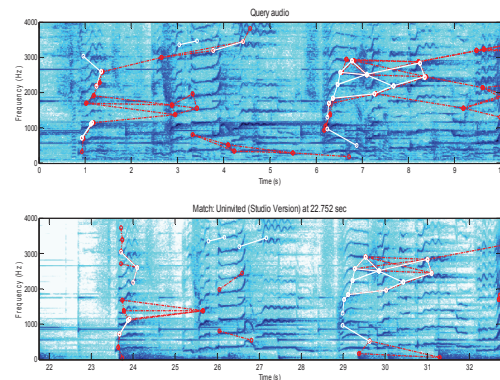


Figure 6: Queried test recording correctly matched against the studio version of a song and not the live version by the same artist

5 NOISE FINGERPRINTING

In this section we discuss the extension of the music audio fingerprinting and alignment approach to the more challenging problem of noise cancellation. We address the question, whether it is possible to identify

and time-align another recording made in the acoustic environment and to use it as a reference signal to cancel out the interference from another signal.

A possible scenario where this would occur is a surveillance situation where there are two microphones within a recording environment. Let us assume that one of the microphones is next to the targets of interest, and the other is closer to the main noise sources in the room. Note that these noises need not necessarily be music or a television broadcast. They could also be from other noise sources like machinery, other speakers, children or simple everyday background noise common to crowded environments like a pub, café or certain meeting rooms. We observe more of the target speech recorded using the microphone that was closer to the target speaker and more of the background noise on the microphone that was farther away from the target. The microphone closer to the target will also pick up a lot of the background noise in the environment, and the microphone farther away will also pick up the same background noise, as well as other sounds that are closer to it.

Now if both the microphones are connected to the same recorder and are simultaneously recording to two separate channels, the problem of reference cancellation is relatively straightforward as both the channels will already be time-aligned, with only the acoustical differences as described above. The more challenging, and more commonly encountered, problem is that of having two microphones in a room that are independently connected to two separate recorders.

These contemporaneous (but not simultaneously clocked) recordings pose a challenging problem to the reference cancellation approach. As in the case of music reference cancellation, the first problem is to time-align the two recordings and the second is to perform the reference cancellation robustly. As the noise recordings may not contain a sufficient number of well-defined spectral peaks, it will be more difficult to obtain a time alignment, and the number of hashes that match across the query and the reference audio will be low. However, even a small number of landmark hashes will probably be sufficient to accurately time align the two recordings. Using these recordings, it is possible to then apply reference cancellation to reduce the effect of the noise.

5.1 Scenario 1: Two independent recordings using two smartphones in the same acoustic environment

Two mobile phones, namely an iPhone 4S and an iPhone 3GS, were used to record a conversation between two speakers in the same acoustic environment.

Since these are two independent devices with their own clocks and digitization hardware, they were not synchronized to each other in any way. Both moving (phones held by the speakers) and stationary recordings (where the phones were placed down) were made. A section of one of the files was used as the query audio, and the recording using the other device was used as the reference file. The landmarking algorithm was applied to both of the files (Figure 7).

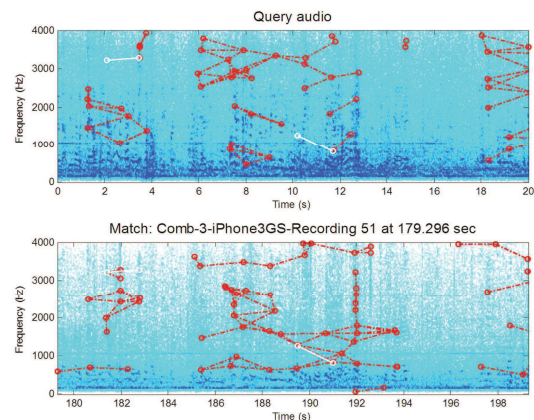


Figure 7: Queried test audio from one recording device (iPhone 4S) matched and time-aligned against a reference recording from another device (iPhone 3GS)

It can be observed that, although the number of matching hashes is significantly less than in the case of the music recording, the two recordings are correctly time aligned.

5.2 Scenario 2: Two fixed microphones in the same acoustic environment

In order to simulate two recordings made from different microphones, we used files from a recording set-up where the two microphones were approximately 4 inches apart. These microphones, placed at different locations in the test environment, were used to feed a 24-track recorder, with each audio signal recorded as a separate wave file, driven by the same sample clock. To simulate separate recording devices, the selected files for the test were trimmed and segmented to have different starting times and lengths, such that although they would overlap in the area of interest, they did not start or end at the same time.

The test recording environment was garage with a concrete floor and hard walls. The basic dimensions of the room were 23 ft. by 24 ft. with a ceiling height of 133 inches. The chamber distinctly had severe acoustic

reverberation, and this reverb is clearly audible in the test recordings. During the recordings, there were no automobiles in the garage, but there were a number of items (e.g., hardware tools, golf clubs, cardboard boxes, etc.) stored along the edges of the walls and the floor. These items do not noticeably affect the reverberation. The garage was located in a suburban location, and there are occasional soft sounds of traffic, nature, and children in the background of the recordings.

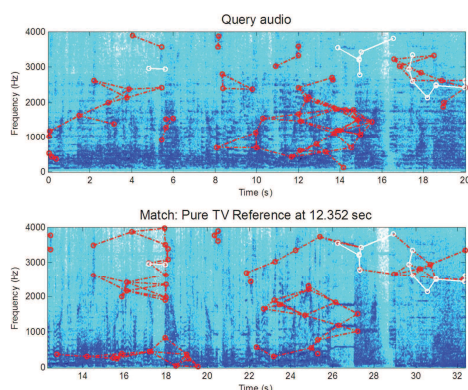


Figure 8: Matching from two microphones in the same recording environment (television reference)

In this experiment the interfering noise was a television broadcast, with two people in the room talking. As previously, while there were a relatively small number of matching hashes as compared with the music landmarking experiments, there were sufficient matches to time-align the two files correctly. The spectrograms of the original (query) audio, the reference channel containing the television interference recorded in the same room, and the resulting speech are shown below.

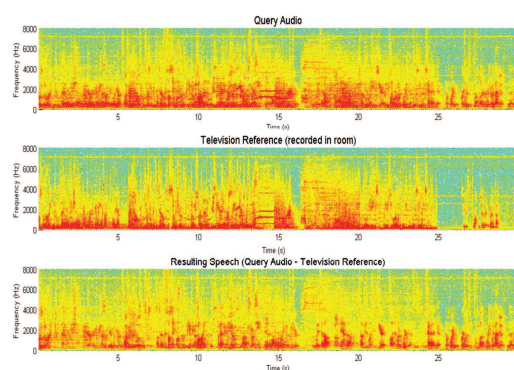


Figure 9: Spectrograms of the query audio, subtracted identified reference microphone signal and the result of noise signal cancellation

5.3 Scenario 3: White noise interference

As a rather extreme test of this approach we used white noise as the interfering source. It is exceedingly difficult to find any distinctive spectral peaks within a white noise recording; however, while the number of matching hashes was significantly less than we observed with either music or regular noise, we were able to identify a very small number of matching hashes that were sufficient to allow time-alignment. Subsequently, the reference cancellation applied using this time-alignment showed significant improvement in intelligibility.

This experiment was then reproduced using source noise files from the CD played back originally. Again, it was possible to correctly identify the appropriate audio track. As shown in the Figure 10 the reference cancellation algorithm is successful in reducing the matched noise and bringing the speech to the forefront.

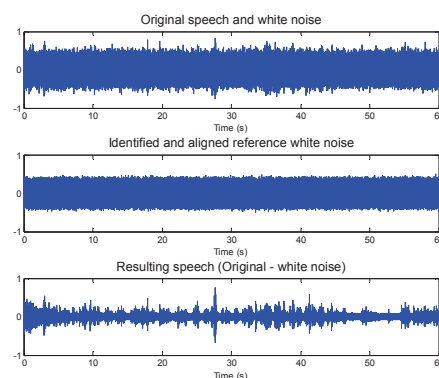


Figure 10: Identification, alignment and reference cancellation of a white noise signal

6 LIMITATIONS

The limitations of this method are that it is not directly applicable to badly clipped, pre-filtered, or heavily compressed recordings, or to recordings where there is a dynamic 'drift', or variable stretch, between the recordings, such as would occur with analog recordings due to wow and flutter, varying tape tension, and motors gradually slowing as the battery voltage drops. Also, what is ultimately extracted may still not be of sufficient quality for recovering usable speech product, due to limited resolution, non-linear distortion, and other possible issues with the underlying signal.

But even with uncompressed digital recordings, timing drift remains a problem, particularly with older technology electromechanical digital devices such as CD players and DAT machines, where the sample clock oscillator can often be off the nominal frequency setting

by several percentage points. With today's direct file playback from personal hand-held music devices, computers, etc., the sample clock on the device and/or soundcard can also be slightly off-nominal, though usually by less than 1%. Although such a small error seems entirely acceptable from a device design standpoint (the music playback won't be noticeably fast or slow), it does pose serious problems to the alignment; for example, a 1% off-nominal sample clock results in a drift of ± 1.2 seconds over a two-minute music track, which is beyond the ability of a reasonably-sized Reference signal canceller to compensate. At present, our land-marking approach does not explicitly account for this drift; however, we believe that further development will allow us to identify multiple alignment points within the two recordings (in the case of music, at least one point near the beginning of the song and one near the end), and then utilize the calculated drift to precisely resample the reference recording to match the speed of the primary.

An interim solution (which, though tedious, has been successfully applied) is to manually pre-process the identified reference track with a software editor, by first determining how short or long it is, relative to the primary over the length of the song, and then calculating and applying the correct stretch factor to match the length of the primary. If the levels between the tracks are not similar, it is also useful to normalize the reference track to maximize the signal level prior to reference cancelling.

7 CONCLUSIONS

In conclusion, our described approach could potentially benefit forensic audio enhancement and transcription by significantly improving the intelligibility of the underlying speech in digital recordings obtained in noisy environments, where a separate digital recording of the interfering audio may be available. Existing LMS-based two-channel reference cancellation approaches can be applied to robustly cancel the interfering audio and leave the speech of target speakers largely intact. A future tool modelled on this landmark-based audio fingerprinting technique could be used both to correctly identify the interfering material, and to quickly and accurately align the reference recording with the primary speech recording.

REFERENCES

- [1] Avery Wang "An Industrial-Strength Audio Search Algorithm", *Proc. 2003 ISMIR International Symposium on Music Information Retrieval*, Baltimore, MD, Oct. 2003.
- [2] J. Benesty, D. Morgan and M. Sondhi, (1997) "A better understanding and an improved solution to the problems of stereophonic acoustic echo cancellation", *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 97, 303
- [3] D. P. W. Ellis. (2009) Robust Landmark-Based Audio Fingerprinting.
<http://labrosa.ee.columbia.edu/matlab/fingerprint>
- [4] Alexander and O. Forth (2011) "No, thank you, for the music': An application of audio fingerprinting and automatic music signal cancellation for forensic audio enhancement", *International Association of Forensic Phonetics and Acoustics Conference 2011*, Vienna, Austria, July 2011