SOUND-BASED CLASSIFICATION OF OBJECTS USING A ROBUST FINGERPRINTING APPROACH

F. Antonacci, L. Gerosa, A. Sarti, S. Tubaro, G. Valenzise

Dipartimento di Elettronica ed Informazione, Politecnico di Milano
P.zza Leonardo da Vinci, 20133, Milano, Italy
email: antonacci/sarti/tubaro@.elet.polimi.it

ABSTRACT

Tangible Acoustic Interfaces (TAIs) are interaction devices that are able to localize the interaction point on a solid surface. Their advantages over traditional interaction devices (touch screens, touch pads, etc.) is in the fact that actual acoustic (vibrational) signals are acquired by contact sensors. This opens the way to interaction classification and recognition.

With this application in mind, this paper approaches the problem of classifying the interaction object from the acquired sounds. We focus on continuous interaction noise, which we classify through a "fingerprinting" approach: features are extracted from the acquired signals and matched against pre-computed features. More sophisticated solutions can be devised for the problem of the classification of noise-like sounds but our approach has the advantage of being computationally simple and can be profitably implemented in real-time.

1. INTRODUCTION

Many research efforts are currently devoted to the development of novel and intuitive devices for Human Machine Interaction (HMI). One of the main problems related to the development of such devices is that active transducers need to be distributed all over the interaction area, with very high production costs. A possible solution to this problem is the use of a limited number of passive acoustic transducers disposed in specific points of the surface and sensitive to events far from their locations (like sounds produced by scratches on the board). In this context, solutions studied in Tai-Chi project showed that an accurate and real-time localization of impulsive and continuous touch of different objects on passive surfaces (such as plexiglass and medium density fiber boards) is possible through Time Differences of Arrival analysis (see [1]). The only requirement of such a system is that the interaction between the object and the board produces a noticeable sound. On the other side one of the advantages of active transducers systems is that we can classify in a relatively simple way the type of object interacting on the sensitive area, thus enabling the system to behave in different ways according to the specific class of object being used. Let us consider, for example, a SmartBoard system used as a blackboard. One may want, for example, to erase the blackboard when a brush is used as interaction tool. The SmartBoard has active sensors under the touchable board. The dimension of the object interacting with the surface informs us of its genre, while assessing the type of interacting object just by using some transducers far from the event could result in a difficult problem to be solved.

This paper concerns with the object classification based on a fingerprinting approach. In the past few years many research efforts have been devoted to classify different types of sounds based merely on their content and without any additional metainformation. The general architecture of fingerprinting frameworks is well described in [2]. Speaking in general terms, a fingerprinting system envisions the extraction of significative features from the sound to be classified and matches the features against a set of pre-computed ones. The main disadvantage of this approach is that the extraction of the features is computationally demanding. Our system has to work on a DSP hardware engaged for the most part of its capabilities with the localization task. For the above presented reasons, we will limit our attention on simple sound classification tools. We will not consider in our discussion more sophisticated solutions (for example based on neural networks or more sophisticated classification framework), because of their computational cost. We will validate the technique presented in this paper with experimental results, showing that different classes of objects can be effectively distinguished in most cases even in a not soundproof context. The rest of the paper is organized as follows: Section 2 will discuss advantages and disadvantages of some fingerprinting algorithms dealing with similar problems. Section 3 will show how standard signal processing tools can be effectively used in our approach. Section 4 will discuss the experimental framework we have used in our system. Section 5 finally summarizes the work and the results.

2. BACKGROUND ON SIMILAR FINGERPRINTING APPROACHES

In the context of digital audio fingerprinting several solutions have been presented in the past few years. Most of them are based on a windowing of the signal as a pre-processing step. In order to reformulate with a common notation some algorithms presented in literature, let define with $x(n)$ the signal and with $x_w(n,l)$ the windowed version of the same signal:

$$x_w(n,l) = x(n-lR)w(n) \quad \text{for} \quad n = 0,1,...,N, \qquad (1)$$

where l denotes a time index, N is the window size and R determines the overlapping between $x_w(n,l)$ and $x_w(n,l+1)$ $(1 \leq R \leq N)$. In fingerprinting applications, Hanning window is generally used and two successive frames overlap for the most part of their length (an overlap factor of 31/32 is a common choice). Using the windowed signals we can compute the STFT of $x(n)$, denoted in the following as $X(f_k,l)$, where, as usual, f_k is the frequency index $(1 \leq k \leq K)$ and l is the time index. In [3] a simple but effective approach for fingerprinting has been presented. The frequency bins of $X(f_k,l)$ are grouped into equally spaced sub-bands, obtaining a low-resolution version of the spectrogram. The sub-band energies

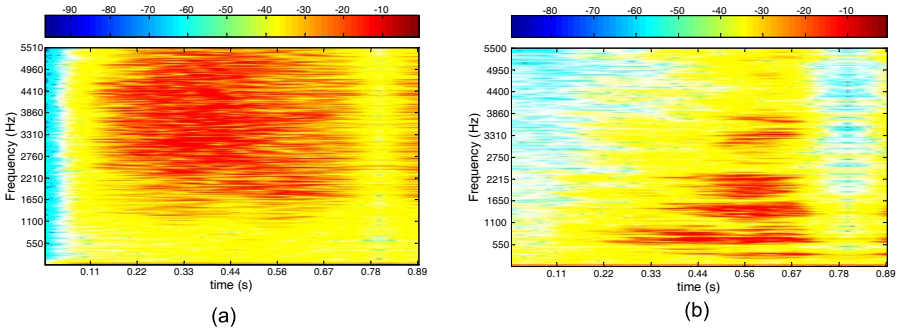

Figure 1: STFTs of the scratch of a sponge (a) a wood stick (b) over a wood board (dB scale).

are obtained as follows:

$$E(m,l) = \sum_{k=mN_f}^{(m+1)N_f-1} |X(f_k,l)|^2, \qquad (2)$$

where m is the sub-band index and N_f is the number of frequency bins in each sub-band.
When the excerpt is corrupted by noise, the distorted and original versions of the same excerpt can be instant by instant different one from the other, but the temporal trends are probably similar, thus revealing the same origin. In order to achieve robustness of the classification system against intentional or non-intentional corruptions of the signal, the approach presented in [3] computes the derivative of $E(m,l)$ with respect to frequency:

$$E_f(m,l) = E(m,l+1) - E(m,l).$$

The time-derivative of $E_f(m,l)$ is computed:

$$E_{ft}(m,l) = E_f(m+1,l) - E_f(m,l).$$

The decision is taken using the 2 levels quantized version $E_{ft}(m,l)$, denoted as $E^q_{ft}(m,l)$.
The feature $E^q_{ft}(m,l)$ is matched against the analogous pre-computed features. The distance between the extracted and the pre-computed features is computed with the Hamming operator.
Even though the approach depicted in [3] is effective, it is not suitable in the case of object classification: in fact the trend of $E_f(m,l)$ is not distinctive of the specific object but of the trajectory and the strength of the specific touch. On the other hand one may think that using a quantized version of $E_f(m,l)$ instead of $E_{ft}(m,l)$ could be the solution of the problem, but preliminary tests showed that the decision made upon a 2-levels quantized version of $E_f(m,l)$ is not satisfactory for our purposes.
In [4] a different approach is presented. Here, once $X(f_k,l)$ is computed, the mean and the standard deviation $M(l)$ and $S(l)$ of each time-frame are computed. In [4] a 3-levels quantized version of the STFT is used:

$$X^{(q)}(f_k,l) = \begin{cases} 0 & \text{if} \quad |X(f_k,l)| \le M(l) \\ 1 & \text{if} \quad M(l) < |X(f_k,l)| \le M(l)+S(l) \\ 2 & \text{if} \quad |X(f_k,l)| > M(l)+S(l) \end{cases} \qquad (3)$$

The best match of $X^{(q)}(f_k,l)$ against the set of pre-computed fingerprints in the database is the decision. Even though simple, this approach results to be suitable, after some modifications, for our purposes.

3. PROPOSED SOLUTION

3.1 Data frequency analysis and noise reduction

The solution proposed in this paper is based on the work of Richly ([4]). In spite of the fact that the above mentioned technique has been developed for musical signals, we treat noise-like signals. We will discuss with examples at hands which modifications we brought to the original system.
In order to be usable, a classification system must preserve the following properties:

- Sounds belonging to different classes are mapped onto different fingerprints (inter-class discriminative requirement.
- Sounds belonging to the same class are mapped onto similar fingerprints (intra-class similarity requirement).

Similarity or difference of fingerprints is based on a distance metric. Further details will be provided in Section 3.4.

In Figure 1 the plots of the STFT of the scratches of a wood stick and a sponge over a wood board are depicted. We can appreciate that the two instruments "sound" different, thus the inter-class discriminative requirement of the system is preserved. At the same time, however, we can observe that different time intervals of the STFT are characterized by different Fourier Transform, which means that the intra-class similarity requirement is not preserved. The following paragraphs will illustrate how to achieve the latter requirement.

Let us call the collection of M successive frames of the STFT with $\mathbf{X}(l)$, with dimensions $K \times M$. The SVD of $\mathbf{X}(l)$ is recalled for convenience in equation (4).

$$\mathbf{X}(l) = \mathbf{U}(l)\mathbf{D}(l)\mathbf{V}^T(l). \qquad (4)$$

The noise-reduced version of $\mathbf{X}(l)$ is obtained according to equation (5)

$$\mathbf{X}^{\{r\}}(l) = \sum_{i=1}^{r} \mathbf{D}(l)_{\{i,i\}}\mathbf{U}(l)_{\{\cdot,i\}}\mathbf{V}(l)^T_{\{i,\cdot\}}, \qquad (5)$$

where r is the rank of the SVD, $\mathbf{D}(l)_{\{i,j\}}$ is ij-th element in $\mathbf{D}(l)$, $\mathbf{U}_{\{\cdot,i\}}(l)$ is the i-th column of $\mathbf{U}(l)$ and, analogously,

$\mathbf{V}(l)_{\{i,\cdot\}}$ is the i-th row of $\mathbf{V}(l)$. In Figure 2 the SVD reduction of the "sponge scratch" is presented using $r = 1$. As expected, we can observe that the spectrum has been smoothed by the low-rank SVD noise reduction. In Section 4 it will be clear that the adoption of the SVD improves the classification performances. In order to reduce the dynamic range

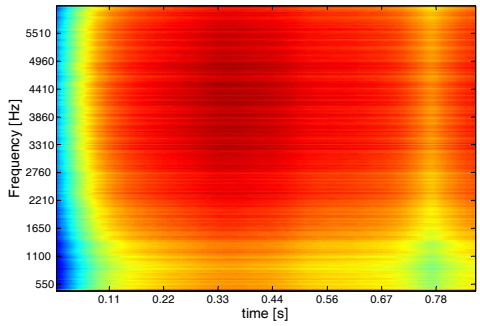

Figure 2: SVD noise reduction of the sponge scratch presented in Figure 1. A decibel scale has been used.

of the signal, we applied also a logarithm compression on $\mathbf{X}^{\{r\}}(l)$.

3.2 Quantization stage

A requirement of sound classification systems is the low computational cost and, as stated above, the intra-class requirement property. In order to fulfill both the requirements we considered different quantization schemes:

1. $M(l)$ and $M(l) + S(l)$ as quantization thresholds, as in equation 3.
2. $M(l)$ as quantization threshold: bins above $M(l)$ are marked with "1", else "0".
3. Uniform quantization with 1,...,8 bits.
4. Lloyd-Max quantization with 2,...,8 levels.

For reasons of space, we will focus our attention on the second quantization scheme. In Figure 3 the 2-level quantization of the sponge scratch is plotted using $M(l)$ as quantization threshold. White and black represent, respectively, the quantization levels '0' and '1'. An interesting side effect of the quantization scheme followed in Figure 3 is its intrinsic adaptivity to the instantaneous dynamic range of the signal. Mean is computed on a frame by frame basis and will adapt itself to the instantaneous dynamic of the signal. The result of the quantization stage is the transformation from the signal $\mathbf{X}^{\{r\}}(l)$ into the new signal $\mathbf{X}^{\{r,q\}}(l)$.

3.3 Training set construction

Once a suitable feature has been singled out, we have to build the library of fingerprints to be used in the following classification step. In order to differentiate the library set from the signal to be classified, each time a signal which is not denoted by the time index l it is part of the fingerprint library set.
The first trivial possibility to build the training set is to use $\mathbf{C}^{(tv)} = \mathbf{X}^{\{r,q\}}$,
where (tv) means time-variant in opposition with time-invariant library set as shown in the next paragraphs.
In Figure 3 we cannot appreciate any noticeable variation of

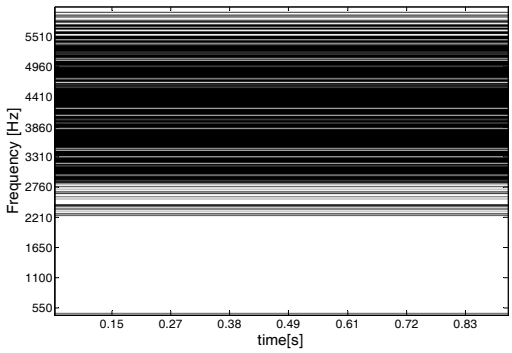

Figure 3: 2 level quantization of the signal in Figure 2.

the fingerprint along the time axis, thus the SVD and quantization stages enable us to preserve the intra-class similarity requirement. This observation suggests us to neglect the variations of $\mathbf{X}^{\{r,q\}}$ across different frames for the construction of the fingerprint library. Using the same notation of equation (5), the transformation of $\mathbf{X}^{\{r,q\}}$ into the time-invariant feature set $\mathbf{C}^{(ti)}$ is illustrated by the following equation:

$$\mathbf{C}^{(ti)}_{\{i\}} = \mathrm{Mo}(\mathbf{X}^{\{r,q\}}_{\{i,\cdot\}}) \,, \tag{6}$$

where 'Mo' denotes the statistical mode. The training set $\mathbf{C}^{(ti)}$ results in a column vector of K elements. In Section 4 we will show that the classification conducted using $\mathbf{C}^{(tv)}$ instead of $\mathbf{C}^{(ti)}$ is a little more effective, but presents the disadvantage of being computationally demanding.

3.4 Classification

Once we have built the training set, we have to classify unknown signals. The classification results in the computation of the distance between $\mathbf{X}^{\{r,q\}}(l)$ and the feature library of each instrument in the database. A fingerprint match is considered 'correct' if the distance between the test sound and its actual class (the actual instrument by which we know a priori the sound was produced) is the minimum in the distance vector. Two different classification schemes were adopted according to the distance metric used. The first distance metric is applicable to both time-invariant and time-variant feature sets, while the second distance metric can be used only with time invariant feature-sets.

- **Feature set $\mathbf{C}^{(ti)}$ and $\mathbf{C}^{(tv)}$:**Pearson correlation. Let us call with $\mathbf{C}^{(tx)}_h$ the time variant or time invariant feature set of h-th instrument in the database. The distance computed with the Pearson correlation between the feature $\mathbf{X}^{\{r,q\}}(l)$ and $\mathbf{C}^{(tx)}_h$ is:

$$\mathrm{D}_P(\mathbf{C}^{(tx)}_h, \mathbf{X}^{\{r,q\}}(l)) = 1 - \frac{\mathcal{Z}[\mathbf{X}^{\{r,q\}}(l)]\mathcal{Z}[\mathbf{C}^{(tx)}_h]}{M-1}, \tag{7}$$

where $\mathcal{Z}(\cdot)$ is the z-score function and the feature matrix $\mathbf{X}^{\{r,q\}}(l)$ has dimensions $K \times M$. Due to the computational cost of Pearson correlation distance, in the experimental section we will use $M = 1$.

- **Feature set $\mathbf{C}^{(ti)}$:**Manhattan distance between the training set $\mathbf{C}^{(ti)}$ and $\mathbf{X}^{\{r,q\}}(l)$:

$$\mathrm{D}_M(\mathbf{C}_h^{(ti)}, \mathbf{X}^{\{r,q\}}(l)) = \frac{1}{RMK} \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \left| [\mathbf{X}^{\{r,q\}}(l)]_{\{k,m\}} - [\mathbf{C}_h^{(ti)}]_{\{k\}} \right|, \quad (8)$$

  where R is the dynamic range of the signal. This way the Manhattan distance $\mathrm{D}_M(\mathbf{C}_h^{(ti)}, \mathbf{X}^{\{r,q\}}(l))$ is normalized across the fingerprint size and the dynamic range of the error and lies in the range [0,1].

Pearson correlation is intrinsically bounded in the range $[-1,1]$, so distance is bounded and normalized in the interval $[0,2]$. Pearson distance has the advantage over Manhattan distance of taking into account not only the punctual element-wise distance but also the linear correlation between the fingerprints. The drawback is the increased computational effort.

4. EXPERIMENTS

4.1 Data-set collection

The training set was built using the objects and surfaces in Table 4.1.

A training set of about 6 recorded sounds per instrument was built. Each sound has been sampled at 44100 Hz, 16 bits, mono and saved in PCM Wave Format. The average length of each sound is about 1 sec. The sounds were acquired in a not soundproof environment, then wavelet coefficients of sound records were thresholded using the SURE soft thresholding algorithm to remove the ambient noise [5]. From each spectrogram the most significant part was selected, to remove attacks and decays of each sound. A simple nearest neighbor interpolation is applied to the spectrograms in order to have the same number of frames.

Table 1: Instruments and surfaces used to build the training set

Instrument	Surface
Highlighting pen	Paper
Iron bar	Wood
Ballpoint pen	Paper
Polystyrene	Wood
Brush	Wood
Wooden Spoon	Wood
Sponge	Wood

4.2 Experimental results

Four dimension of analysis have been considered, which are orthogonal to the quantization policies described in Section 3. In particular, the following testbeds have been implemented:

- SVD/noSVD: 3-levels quantization with manhattan distance computed on the whole fingerprint.
- log/ no log compression: a logarithmic compression function can be applied after the spectrogram coefficients have been reduced by the SVD.
- Full/1 frame distance: distance can be computed using all the fingerprint coefficients or taking only the central frame of the quantized spectrogram.
- Manhattan/correlation distance: when distance is computed across one frame only, the distance function can be manhattan or Pearson correlation distance.

To evaluate the quality of the match for each method, the following procedure has been applied. For each element of the test-set, the distance (manhattan or correlation) between the fingerprint and the fingerprints stored in the database has been computed. Different samples of the same instrument show very similar distance behavior, so it is possible to make statistics of each method per instrument rather than per sound record. Therefore, an average distance vector for each instrument, computed as the mean of three test sound records for that instrument is computed for each testbed. A rank of the quality of the match is assigned to each instrument in the testbed. To compute this rank, the vector containing the average distance per instrument is sorted in ascending order and the difference between the second and the first entry of the vector is computed. This absolute rank is then divided by the minimum distance to obtain a per cent relative rank. The main advantage of this representation is that one has an immediate way of detecting false matchings, which have negative ranks.

Using the test procedure illustrated above, Figure 4 was generated. Fingerprint extraction without SVD has a low rate of correct matches ($\sim 28\%$), while in the other tests the correct match rate is 100%. This motivated the use of SVD on the spectrogram to reduce the effect of noise.

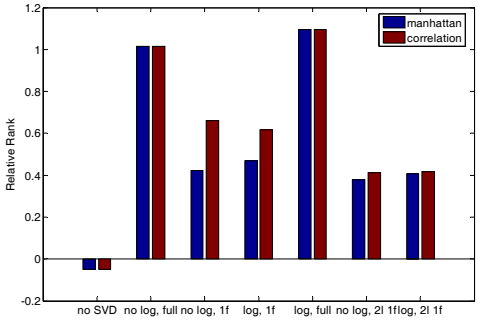

Figure 4: Relative rank for some testbeds using 2 and 3 levels of quantization, using manhattan and correlation 1-frame distance. Full distance is computed using Manhattan distance.

From Figure 4 one can make two considerations. First of all, applying a log compression to SVD result does not change noticeably the matching result. Second, matching carried out computing distance over all the fingerprint matrix is better than matchings performed on one frame only.

Figure 5 shows the rate of successful matches using Lloyd-Max quantization. Here the distances are computed using the central frame of the fingerprint matrix reduced by SVD, without any log compression. The results are compared in the case of manhattan distance and correlation distance. We can appreciate that the quantization that adopt $M(l)$ and $M(l)+S(l)$ as quantization thresholds is more effective than other techniques, even though it has less quantization levels. The reason of this resides in its intrinsic adaptivity to the dynamic range of the signal.

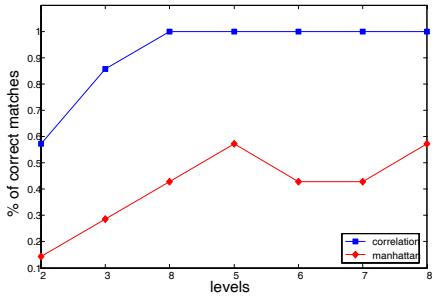

Figure 5: Rate of correct matches for LLoyd-Max quantization with Manhattan and correlation distance functions.

5. CONCLUSIONS

In this paper we have shown a method that allows to assess the object scratching over o solid surface by acquiring a vibrational signal. The technique can be effectively used in the context of Tangible Acoustic Interfaces (TAIs) relying simply on acoustic signals. Results demonstrate that accurate classification can be achieved.

REFERENCES

[1] A. Sarti G. Scarparo G. De Sanctis, D. Rovetta and S. Tubaro. Localization of tactile interactions through tdoa analysis: Geometric vs. inversion-based method. In *Proceedings of 2006 European Signal Processing Conference (EUSIPCO 2006)*, Sept. 2006.

[2] P. Cano, E. Batle, T. Kalker, and J. Haitsma. A review of algorithms for audio fingerprinting. *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 169–173, 2002.

[3] T.Kalker and J.Haitsma. A highly robust audio fingerprinting system. In *Proceedings of ISMIR 2002, 3rd International Conference on Music Information Retrieval*, October 2002.

[4] F.Kovacs G.Richly, L.Varga and G.Hosszu. Short term sound stream characterization for reliable, real-time occurrence monitoring of given sound-prints. In *Proceedings of MEleCon 2000, 10th Mediterranean Electrotechnical Conference*, volume II, May 2000.

[5] D. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.