A robust howling detection algorithm based on a statistical approach

Conference Paper · September 2014

DOI: 10.13140/2.1.3351.9042

CITATION
1

READS
585

4 authors, including:

Alexandre Guerin
Orange Labs
29 PUBLICATIONS **255** CITATIONS
SEE PROFILE

Pascal Scalart
Université de Rennes 1
149 PUBLICATIONS **1,759** CITATIONS
SEE PROFILE

A ROBUST HOWLING DETECTION ALGORITHM BASED ON A STATISTICAL APPROACH

Joachim Flocon-Cholet[⋆] *Julien Faure*[⋆] *Alexandre Guérin*[⋆] *Pascal Scalart*[†]

[⋆] Orange Labs, Lannion - France
joachim.floconcholet@orange.com
[†] INRIA/IRISA, Université de Rennes 1, Rennes - France

ABSTRACT

This paper presents an algorithm for the detection of howlings that arise in audio signals. Our method is based on the combination of two energy-based features and one new feature related to the frequency stability of a howling component. The decision stage, which implies a Support Vector Machine (SVM) model, outputs a decision every 20 ms. The evaluation, carried out on a large database, showed that the algorithm is able to detect both pure and multiple tones howling in a wide range of energy. Furthermore, even on complex signals such as music, the detection is still efficient with very few false alarms.

Index Terms— Acoustic signal detection, Howling detection

1. INTRODUCTION

For the last decade, there has been an increase in the number of people who work in call center environments using headsets or handsets throughout their work day. Correspondingly, there has been an increase in the number of reports where call center staff were exposed to an unexpected and sudden howling. The term howling is generally used by the listener to describe a pure or multiple tone that arises in an audio signal. In [1], the author details the physiological consequences of exposure to howlings, which can cause at least a discomfort, but more generally, a stress, or headaches, even when it appears at low level.

To address the problem of howling detection, methods found in the literature are usually based on a combined spectral and temporal analysis of the audio signal. As described in [2], most algorithms first select frequency components having the largest magnitude which are considered to be candidate howling components. Then, different features are computed on these candidates in order to exhibit the temporal and spectral properties of the howling components, such as, the power ratio of the candidate howling component and the entire spectrum, or the persistence of the same candidate howling component for a certain duration. See [3] for list of the commonly used features. The final decision, i.e. the howling component is present or not, is obtained by comparing the computed features to predefined thresholds [2], [4], [5].

As pointed out in [2], these methods suffer from the use of fixed thresholds which have to be tuned carefully to adjust the trade-off between detection rate and false alarm rate. A low detection rate indicates that some howlings will not be detected and suppressed, which can cause a discomfort for the end-user. On the other hand, a high false alarm rate implies that non-holwing components will be detected and suppressed leading to a signal degradation. Furthermore, the tuning is generally done for a specific use case, i.e. the howling appears on speech signal only or on music signal only. As a consequence, a given method can achieve good results if the analysed signal is speech but might have a high probability of false alarm in presence of music signal. An other limitation is the assumption that a howling is a pure tone while it can be a multiple tone as well. In this case, the detection is not complete.

An alternative method can be found in [6], where the authors detail an algorithm for pure tones detection in audio signals using two features that feed a two-state Hidden Markov Model (HMM). The first feature is the zero-crossing rate and the second one uses the 95th percentile to highlight strong peaks within the [900 Hz - 4 kHz] frequency band. This procedure is well designed for pure tones detection but not for multiple tones detection. Furthermore, a high probability of false alarms might be observed for complex signals, such as music, which contain strong high frequencies.

In order to tackle the problems mentioned above, we present a howling detection algorithm based on a Support Vector Machine (SVM) model using two energy-based features and one novel feature, namely the frequency stability feature. The remaining of the paper is organized as follows: the audio features and the principle of SVM are described in sections 2 and 3 respectively. In section 4, we present an evaluation, carried out on a large database, in order to compare both our method and the solution proposed in [6]. The results show the benefits of our approach where good performance is obtained for different conditions.

2. AUDIO FEATURES

We consider that two properties are relevant to characterize a howling component. First, it can be loud compared to the rest of the signal, or at least, compared to the surrounding frequencies. Second, a howling component is constant over time both in amplitude and frequency. To implement these ideas, we adopt a sub-band approach. At each frame t, the Fourier spectrum is divided in several sub-bands and on each of these sub-bands, the three features are computed.

Based on that, we define three audio features: the Local Spectral Crest Factor, the Global Spectral Crest Factor and the Frequency Stability. Throughout the paper, we will denote X_k^t the amplitude of the frequency bin k of the Fourier spectrum at frame t.

2.1. Local Spectral Crest Factor

We make the assumption that if a howling arises in the signal, we will observe a peak in the frequency sub-bands where the howling component is present. Therefore, in each sub-band, we measure the spectral flatness by computing the ratio between the maximum and the mean energy in the sub-band, as it was proposed in [7]. Following the authors of [7], we will call $S_t(B)$ the Local Spectral Crest Factor (LSCF) at time t, in the frequency band B such as:

$$S_t(B) = \frac{\max_{k \in B} X_k^t}{\frac{1}{K} \sum_{k \in B} X_k^t}, \tag{1}$$

where K is the number of frequency bins within the sub-band B.

2.2. Global Spectral Crest Factor

The LSCF indicates the presence of a dominant peak in a sub-band. However, this peak could be neglected regarding the overall spectrum energy. In order to compare a peak in a sub-band to the entire spectrum, we define $G_t(B)$, the Global Spectral Crest Factor (GSCF) at time t in the frequency band B, as:

$$G_t(B) = \frac{\max_{k \in B} X_k^t}{\frac{1}{N} \sum_{k=1}^{N} X_k^t}, \tag{2}$$

where N is the number of frequency bins of the entire spectrum.

2.3. Frequency Stability

A howling component has the property of being relatively constant over time, both in frequency and amplitude. This assumption has been used in previous work as stated in [3], with the *Interframe Peak Magnitude Persistence* (IPMP). This feature is based on counting the number of the past consecutive frames a frequency bin X_k has been a candidate howling component.

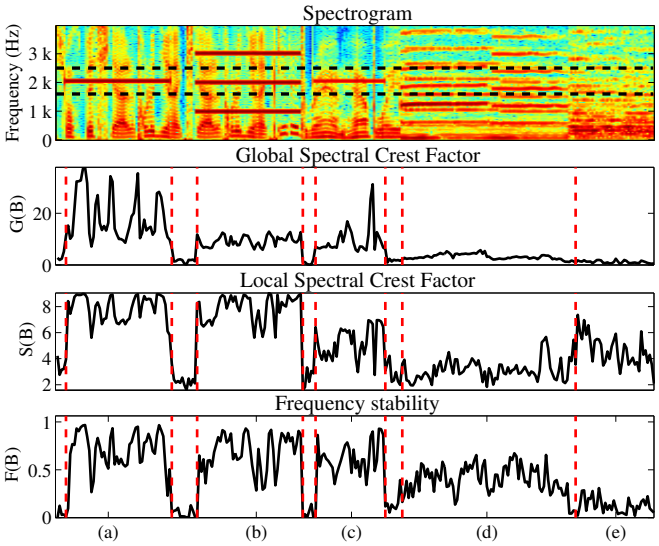

Fig. 1. Behaviour of the three features on five different situations: (a) loud pure tone, (b) loud multiple tone, (c) low-level pure tone, (d) music with horns and (e) orchestral music.

However, a simple measure of persistence might cause false alarm on musical signals where long notes are played. Hence, we propose the Frequency Stability feature to emphasize small variations of amplitude between two consecutive bins X_k^t and X_k^{t-1} together with high X_k^t. For that purpose, we define $F_t(B)$, the Frequency Stability at time t, in the sub-band B as:

$$F_t(B) = \frac{1}{K} \sum_{k \in B} \exp\left(-\Delta_k^{t,t-1}\right) \frac{X_k^t}{\overline{X_B}}, \tag{3}$$

where $\Delta_k^{t,t-1} = \frac{|X_k^t - X_k^{t-1}|}{X_k^t}$ is the spectral variation between frames t and $t-1$, and $\overline{X_B} = \frac{1}{K} \sum_{k \in B} X_k^t$ is the mean energy within the sub-band B. $F(B)$ can be viewed as a normalized sub-band spectrum $\left(\frac{X_k^t}{\overline{X_B}}\right)$ weighted by a frequency stability term $\left(\exp\left(-\Delta_k^{t,t-1}\right)\right)$.

We note that the three features are computed using the current frame, for the LSCF and GSCF, or both the current and the previous frames for the Frequency stability. These short-term features are thus suitable for low-delay applications.

2.4. Case study

Figure 1 illustrates the behaviour of the above features for an audio sample containing five cases: (a) a loud pure tone howling, (b) a loud multiple tone howling, (c) a lower-level pure tone howling and two musical excerpts composed of (d) horns and (e) orchestral music. The features are computed for the frequency band delimited by the dashed lines on the spectrogram.

As it can be observed, the GSCF gives high values for the loud pure tone (a) but collapses for the cases of multiple tone and low-level pure tone. In contrast, the LSCF highlights well both loud pure tone and multiple tone. However, this feature also collapses on (c) due to the low level of the pure tone regarding the energy within this frequency band. Furthermore, on the orchestral music part (e), the LSCF tends to produce false alarms due to isolated harmonics.

As the Frequency Stability feature emphasizes the steadiness over time of a frequency component, it shows well the presence of the pure and multiple tone, even when the tone has a low energy. However, on the musical part where long notes are played (d), this feature might cause false alarms.

These examples show that each feature is able to characterize a specific situation; the combination of these three features should lead to a robust detection of pure and multiple tones.

3. CLASSIFICATION

3.1. Support Vector Machines (SVMs)

The decision step involves the use of a classifier and we focused our attention on SVMs which have been very popular in the recent years and have proven to be effective in various classification tasks [8]. The idea of SVMs is to find a decision boundary between two classes by mapping the training sample vectors onto a higher dimensional space and then determining an optimal separating hyperplane [9].

Let us consider a training data set composed of M samples of D-dimensional feature vectors $\boldsymbol{x}_i$, associated with a class label y_i, such that: $\boldsymbol{x}_i \in \mathcal{R}^D$, $y_i \in \{-1, +1\}$ and $i = 1, ..., M$. The algorithm searches for an optimal hyperplane that separates the data with a maximal margin by solving the optimization problem:

$$\text{minimize} \qquad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum \xi_i \tag{4}$$

$$\text{subject to} \qquad y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \tag{5}$$

where $\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$ denotes the dot product between normal vector to the hyperplane $\boldsymbol{w}$ and a feature vector $\boldsymbol{x}_i$, b the offset and the ξ_i are the slack variables to handle the case where data are not separable, i.e some points can be on the wrong side of the hyperplane. The penalty parameter C controls the trade-off between classification errors and maximum margin.

Furthermore, when a linear decision boundary is not sufficient, we can project the data onto a higher dimensional space using a transformation function $\Phi(\boldsymbol{x})$ where it is possible to define a linear hyperplane. Since the solution is formulated as a dot product, we can define the kernel function such that $k(\boldsymbol{x}, \boldsymbol{x}_i) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}_i) \rangle$. Solving this problem, we obtain the following decision function:

$$f(\boldsymbol{x}) = \left(\sum_{i=1}^{M} \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b \right), \tag{6}$$

where the α_i are the Lagrange multipliers associated with the optimization problem. The label of a new input is obtained by looking at the sign of $f(\boldsymbol{x})$.

We choose the Radial Basis Function (RBF) kernel: $k(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-\gamma||\boldsymbol{x} - \boldsymbol{x}_i||^2)$, which gave the best results.

3.2. Temporal integration

To prevent some discontinuities in the sequence of the decisions, we use a HMM with two states (*Howling*, *No howling*), using the output of the SVM. We first derive SVM probabilities by mapping $f(\boldsymbol{x})$ on a sigmoid function, as in [10]:

$$P(y = +1|f(\boldsymbol{x})) = \frac{1}{1 + \exp(Af(\boldsymbol{x}) + B)}, \tag{7}$$

where A and B are derived from the distribution of $f(\boldsymbol{x})$ in the training step. The observation distributions of the HMM are obtained by modeling $P(y = +1|f(\boldsymbol{x}))$ with 6 Gaussians, fitted with the Expectation/Maximization algorithm. For each frame, the state is deduced with the Viterbi algorithm using observations of the current frame and the states of the previous frames. Thus, no extra delay is added.

4. EVALUATION

4.1. The datasets

To compare our algorithm with the method proposed in [6], we consider three cases: (a) pure tone howlings appearing on speech signal, (b) pure tone howlings appearing on mixed signal (i.e signal composed of speech, music and mix of speech over music) and (c) both, pure and multiple tone howlings, appearing on mixed signal. Each of these conditions corresponds to a test set denoted *PureOnSpeech*, *PureOnMixed* and *MultiOnMixed*, respectively. These sets are created from a database sampled at 8 kHz, which contains speech sentences of dozen of languages and music samples covering a large variety of musical genre such as classical music, jazz, hip hop etc. Each database is normalized to have a mean level of -26 dB FS.

Pure tones are generated randomly with a fundamental frequency ranging from 900 Hz to 4 kHz. For multiple tones we considered square and multiple harmonic signals which were generated in the same way as the pure tones. The level of a howling is drawn randomly within the range of -40 to 0 dB FS. Each howling is one second long.

In order to train our algorithm, we also created a training set similar to *MultiOnMixed*. We ensure that the training set and the different test sets are independent. The different datasets are summarized in table 1, according to their type of signal, type of tone, number of tones (N) and overall duration.

Datasets	Signal	Howling type	N	Duration
PureOnSpeech	Speech	Pure	500	45min
PureOnMixed	Mixed	Pure	45	8min
MultiOnMixed	Mixed	Pure+Multi	45	8min
Training	Mixed	Pure+Multi	76	4min

Table 1. Composition of the datasets.

4.2. Protocol

The audio signal is first segmented into frames of 20 ms. For each frame, a Fast Fourier Transform (FFT) is computed using 256 bins. The spectrum is then decomposed into 12 rectangular bands spaced on a mel scale with 50% overlap. The three features, i.e. the LSCF, the GSCF and the Frequency Stability, are computed for each frequency band and each frequency band is used as a dimension of the feature vector. Each frame is thus associated to a 36 dimensions feature vector.

The SVM with two classes (*Howling*, *No howling*), is trained on the training set. A grid-search procedure is performed to find the best SVM's parameters, C and γ, which are the penalty and kernel parameters, respectively, by ten folds cross-validation on the training set. We used the *LibSVM* library [11] as implementation of the SVM. The HMM is trained using the outputs of the SVM on the training set.

The evaluation is done by comparing the predicted label to the annotated label for each frame.

4.3. Results

The results of the evaluation are presented in figure 2. For each test set, we give the detection rate and the false alarm rate for our proposition and the state of the art method [6].

On the *PureOnSpeech* test set, both algorithms show good performance but our algorithm gives higher accuracy and reduces the false alarm rate by a factor 5.

For the *PureOnMixed* test set, the detection rate drops for both algorithms which indicates the difficulty of detecting pure tones on complex signals. The important result for this test set is the increase of the false alarm rate for the state of the art algorithm (from 2.6 % on *PureOnSpeech* to 17.6 % on *PureOnMixed*) while our method keeps the false alarm rate low (1.6 %). This highlights that our features mainly describe the characteristics of a pure or multiple tone and don't identify harmonics in a music signal as howlings.

For the last test set, *MultiOnMixed*, our algorithm still shows decent performance proving that it is able to detect both pure and multiple tones, even in the presence of mixed signals. On the contrary, we can see that the state of the art method is solely designed for pure tone howling detection.

Figure 3 depicts the detection rate for the *PureOnMixed* test set according to the Signal to Tone ratio. This ratio is computed using the frame energy of the original signal and the frame energy of the howling. As it can be observed, even

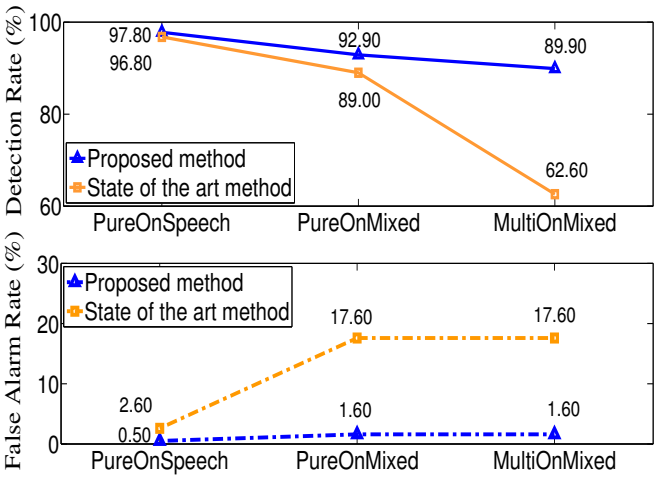

Fig. 2. Detection rate and False alarm rate for the proposed method and the state of the state of the art method on the three datasets.

when a pure tone has a lower energy compared to the speech or mixed signal, a fair detection rate is achieved with the proposed method, and outperforms the state of the art algorithm.

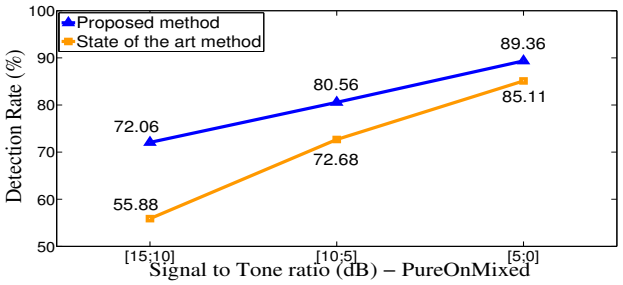

Fig. 3. Comparison of the detection rate for different Signal to Tone ratios on the *PureOnMixed* test set.

5. CONCLUSION

We proposed a robust low-delay algorithm for howling detection using a statistical approach. This method relies on two energy-based features and one novel features related to the frequency stability of a howling component, combined with a SVM model. The evaluation performed on a large database showed that, unlike previous methods, our algorithm handle the detection of both pure and multiple tone howlings even at low level, which proves that our features combined with the SVM model truly capture howling components. Furthermore, the proposed method can be used in different situation since its works properly on speech, music and mixed signals, with a very low false alarm rate.

6. REFERENCES

[1] M. Westcott, "Acoustic shock injury (asi)," *Acta Oto-Laryngologica*, vol. 126, no. S556, pp. 54–58, 2006.

[2] T. van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: state of the art and future challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288–327, Feb. 2011.

[3] T. van Waterschoot and M. Moonen, "Comparative evaluation of howling detection criteria in notch-filter-based howling suppression," *J. Audio Eng. Soc.*, vol. 58, no. 11, pp. 923–940, Nov. 2010.

[4] G. Choy, D. Hermann, R. L. Brennan, T. Schneider, H. Sheikhzadeh, and E. Cornu, "Subband-based acoustic shock limiting algorithm on a low-resource dsp system.," in *INTERSPEECH*, 2003.

[5] T. Soltani, D. Hermann, E. Cornu, H. Sheikhzadeh, and R. L Brennan, "An acoustic shock limiting algorithm using time and frequency domain speech features.," in *INTERSPEECH*, 2004.

[6] J. Faure, A. Guérin, and C. Marro, "Method and device for detecting acoustic shocks," Jan. 6 2012, WO Patent 2,012,001,261.

[7] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," 2004.

[8] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1885–1888.

[9] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond (Adaptive computation and machine learning)*, The MIT Press, 2001.

[10] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[11] C-C. Chang and C-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.