Human Whistle Detection and Frequency Estimation

Mikael Nilsson, Josef Ström Bartůněk, Jörgen Nordberg, and Ingvar Claesson
Department of Signal Processing, School of Engineering, Blekinge Institute of Technology
Box 520, SE-372 25 Ronneby, Sweden
E-mails: mkn@bth.se, jsb@bth.se, jno@bth.se, icl@bth.se

Abstract

Human whistle could be a way to perform activation of different kind of devices, for example turn on and off a light in a smart room. Therefore, in this paper a human whistle detection and frequency estimation system is presented. Further, an investigation of human whistling and a robust non-linear feature extraction is presented. A system for robust performance due to sensor change and various noise situations is proposed using these features. Experiments in various noise situations are conducted.

1 Introduction

Most humans have the ability to whistle. Human whistling is typically single frequency dominated signals with a distinct characterization, although harmonics might occur. Whistling is produced by means of a constant airflow from the lungs. The air is moderated by the tongue, lips, teeth or fingers to create turbulence, and the mouth acts as a resonant chamber. Whistling can be considered as a simple way of communication between humans, typically to bring attention.

General whistling, human or non human, is a communication mean in various situations; for example dolphins whistle and referees whistle in soccer games [8, 11]. Furthermore, due to the characteristics of whistling, there is a close connection to single tone detection [3, 4]. However, human whistling is an underexposed and a remarkable area of expression, it is raw and direct. The applicability of human whistling detection can be manifold. It can be used as an aid for handicapped to activate alarms, activate lights or other devices in a smart room. Some work has addressed the usefulness of human whistle [1, 7]. However, to the best of the author's knowledge, no detailed analysis or description of a digital detection algorithm can be found for human whistling.

In this paper, the human whistle is investigated by time-frequency analysis. Results from the analysis is used to propose a robust feature extraction scheme for whistle detection. To achieve robustness the non-linear technique called the Successive Mean Quantization Transform (SMQT) [9, 10] is used in this paper. The SMQT has properties that reveal the underlying structure in data; hence it will reduce or remove dissimilarities due to different sensors used. Experiments are conducted on real signals and the system is investigated under noisy situations.

The paper is organized as follows. In the next section the analysis of the human whistling is performed. Section 3 presents a short description of the Successive Mean Quantization Transform (SMQT). Section 4 discusses the proposed feature extraction. Section 5 presents decision rules given the features. In Section 6 experimental results for the proposed whistle detector are highlighted. Conclusions are given in Section 7.

2 Human Whistle Characteristics

Human whistles vary in frequency. Hence it is desired to find a typical frequency range for human whistling. To do so, a database with different people whistling has been created. The database consists of 20 randomly selected test subjects and has shown, from experimental results, to be enough for this initial study. The test subjects were told to whistle melodies and to try to achieve as high and low frequency whistle sound as possible. All these recordings are performed in a noise free environment. Note, that the recorded signals are typically non-stationary, since the signal may contain no whistle or different whistle frequency at different times. The signal potentially containing human whistle is denoted $s(n)$ where n is the time discrete index. The sampling rate used in this paper is chosen to 48 kHz.

The aim with this analysis is to find the lower and upper frequency limits for human whistle. In the analysis of human whistle the Power Spectral Density (PSD) is estimated using Welch's method [6]. The calculation of the mean PSD is done by using a block size of 512 samples, 50% overlap and a Hamming window, see Fig. 1.

From Fig. 1, it can be seen that the human whistle is typ-

978-0-7695-3119-9/08 $25.00 © 2008 IEEE
DOI 10.1109/CISP.2008.415

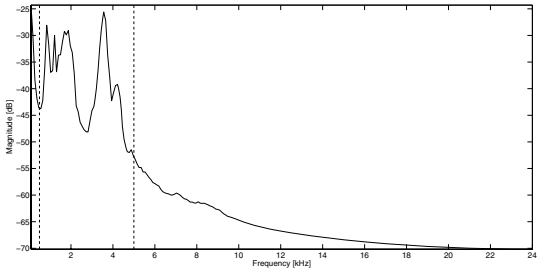

Figure 1. Average Power Spectral Density from whistle database.

ically located in the range of 500-5000 Hz. Of course, some people might exceed these limits, such as trained whistlers. However within these limits it is possible for most people to produce whistles. Even if the signal of interest is typically below 5000 Hz it is still interesting to have information about higher frequencies, since in some signals, i.e. music, whistle-like sounds can occur but information from the higher frequencies can avoid such false detections.

Given these limits two order-100 Hamming window based Finite Impulse Response (FIR) filters are designed [5]; one bandpass filter (H_{bp}) and one bandstop filter (H_{bs}) both using 500-5000 Hz as a pass- and stopband respectively, see Fig. 2.

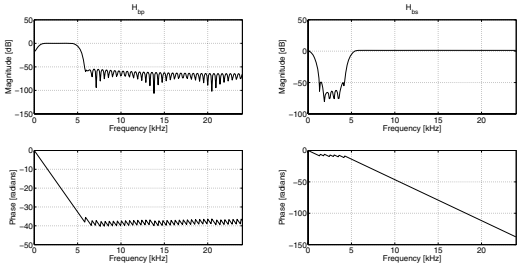

Figure 2. Filters H_{bp} and H_{bs}.

3 Description of the SMQT

Part of the calculation of the feature vectors involves using the Successive Mean Quantization Transform (SMQT) [9, 10]. A short description of the SMQT is given here for convenience. Note that the description use set theory notation.

Let x be a data point and $\mathcal{D}(x)$ be a set of $|\mathcal{D}(x)| = D$ data points. The value of the data point will be denoted $\mathbf{V}(x)$. The SMQT has only one parameter input, the level L. The output set from the transform is denoted $\mathcal{M}(x)$. The transform of level L from $\mathcal{D}(x)$ to $\mathcal{M}(x)$ is denoted

$$\mathrm{SMQT}_L : \mathcal{D}(x) \rightarrow \mathcal{M}(x) \tag{1}$$

The SMQT_L function can be described by a binary tree where the vertices are Mean Quantization Units (MQUs). A MQU consists of three steps, a mean calculation, a quantization and a split of the input set.

The first step of the MQU finds the mean value of the data, denoted $\overline{\mathbf{V}}(x)$, according to

$$\overline{\mathbf{V}}(x) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbf{V}(x) \tag{2}$$

The second step uses the mean to quantize the values of the data points into $\{0, 1\}$. Let a comparison function be defined as

$$\xi\left(\mathbf{V}(y), \overline{\mathbf{V}}(x)\right) = \begin{cases} 1, & if\ \mathbf{V}(y) > \overline{\mathbf{V}}(x) \\ 0, & else \end{cases} \tag{3}$$

where $y \in \mathcal{D}$. Further, let $\bigotimes$ denote concatenation, and then

$$\mathcal{U}(x) = \bigotimes_{y \in \mathcal{D}} \xi\left(\mathbf{V}(y), \overline{\mathbf{V}}(x)\right) \tag{4}$$

is the mean quantized data set. The set $\mathcal{U}(x)$ is the main output from a MQU. The third step splits the input set into two subsets

$$\begin{aligned} \mathcal{D}_0(x) &= \{x \mid \mathbf{V}(x) \leq \overline{\mathbf{V}}(x), \forall x \in \mathcal{D}\} \\ \mathcal{D}_1(x) &= \{x \mid \mathbf{V}(x) > \overline{\mathbf{V}}(x), \forall x \in \mathcal{D}\} \end{aligned} \tag{5}$$

where $\mathcal{D}_0(x)$ propagates left and $\mathcal{D}_1(x)$ right in the binary tree, see Fig. 3.

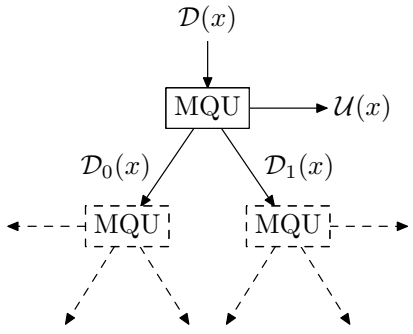

Figure 3. The operation of one Mean Quantization Unit (MQU).

The MQU constitutes the main computing unit for the SMQT. The first level transform, SMQT_1, is based on the

output from a single MQU, where $\mathcal{U}$ is the output set at the root node. The outputs in the binary tree need extended notation. Let the output set from one MQU in the tree be denoted $\mathcal{U}_{(l,n)}$ where $l = 1, 2, \ldots, L$ is the current level and $n = 1, 2, \ldots, 2^{(l-1)}$ is the output number for the MQU at level l, see Fig. 4. Weighting of the values of the data points

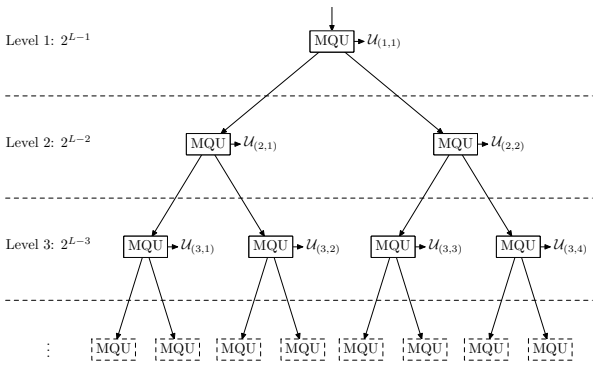

Figure 4. The Successive Mean Quantization Transform (SMQT) as a binary tree of Mean Quantization Units (MQUs).

in the $\mathcal{U}_{(l,n)}$ sets are performed and the final SMQT_L is found by adding the results. The weighting is performed by 2^{L-l} at each level l. Hence, the result for the SMQT_L can be found as

$$\begin{aligned}\mathcal{M}(x) \quad = \quad & \{x \mid \mathbf{V}(x) = \textstyle\sum_{l=1}^{L} \sum_{n=1}^{2^{l-1}} \mathbf{V}\left(u_{(l,n)}\right) \cdot 2^{L-l}, \\ & \forall x \in \mathcal{M}, \forall u_{(l,n)} \in \mathcal{U}_{(l,n)}\}\end{aligned} \tag{6}$$

As a consequence of this weighing the number of quantization levels, denoted Q_L, for a structure of level L will be $Q_L = 2^L$.

4 Calculation of Feature Vectors

Given the human whistle characteristics, we extract robust features for whistle detection. To extract these features, the signal $s(n)$ will undergo the steps outlined in Fig. 5.

The first step divides $s(n)$ into non-overlapping blocks of size 512. To extract robust features on these blocks a Successive Mean Quantization Transform (SMQT) of level L (SMQT_L) is applied to each block, as described in the previous section. In this paper $L = 8$ is used at all times. The SMQT will make the features robust to various sensor changes. It reduces or removes the effect of different microphones, different dynamic range, bias shift and gain shift. The output from the SMQT is in the range $[0 \ldots 255]$

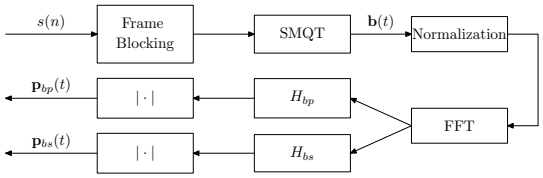

Figure 5. The steps from signal $s(n)$ to the feature vectors $\mathbf{p}_{bp}(t)$ and $\mathbf{p}_{bs}(t)$

(since $L = 8$). Normalization of the result from the SMQT is performed by

$$\frac{\mathbf{b}(t) - 2^{L-1}}{2^{L-1}} \tag{7}$$

where $\mathbf{b}(t)$ is the SMQT result from block t. This normalization will ensures that the values in the result will be guarantied to be in the range $[-1, 1]$.

A 512-point Fast Fourier Transform (FFT) is applied to the normalized result. Further the bandlimiting filters $H_{bp}(k)$ and $H_{bs}(k)$ are applied on the FFT output, where k denotes the discrete frequency. Finally, the absolute value of the filtered results yields the feature vectors $\mathbf{p}_{bp}(t)$ and $\mathbf{p}_{bs}(t)$. A truncation to 256 values is performed due to symmetry from the FFT on real signals. Hence, two feature vectors, $\mathbf{p}_{bp}(t)$ and $\mathbf{p}_{bs}(t)$, of size 256 are found for every block t.

5 Detection and Frequency Estimation

To detect human whistle, the feature vectors $\mathbf{p}_{bp}(t)$ and $\mathbf{p}_{bs}(t)$ will be examined, see Fig. 5. The following observations are the motivation for the design of the robust whistle detector:

I The largest value in $\mathbf{p}_{bp}(t)$ should typically be larger than the mean of $\mathbf{p}_{bs}(t)$ in the presence of whistle.

II In presence of whistle $\mathbf{p}_{bp}(t)$ has typically a few very dominant values.

III If I and II are not fulfilled there is no distinct whistle tone or it is totally drowned in noise.

IV To avoid noisy single block detections a smoothing of the detection results should be performed. Such noisy detections could typically occur if music is present in the signal.

In order to implement the above points, it is necessary to extract the following quantities for each block t. The maximum value, $\max \mathbf{p}_{bp}(t)$, as well as the vector index of the

maximum value, $\arg\max \mathbf{p}_{bp}(t)$, are found from $\mathbf{p}_{bp}(t)$. Note that the index of the maximum value can be considered to be an estimate of the discrete fundamental frequency index for the whistle, provided whistle sound is present. Further, the mean value, $\text{mean}\ \mathbf{p}_{bs}(t)$, is calculated from $\mathbf{p}_{bs}(t)$. A ratio is found as

$$\gamma(t) = \frac{\max \mathbf{p}_{bp}(t)}{\text{mean}\ \mathbf{p}_{bs}(t) + 1}. \tag{8}$$

For non-whistle blocks $\gamma(t)$ will have lower values than if whistle is present. A threshold on $\gamma(t)$ will be the first requirement in the decision if whistle is present or not

$$a(t) = \begin{cases} 1, \text{ if } \gamma(t) > \theta_\gamma \\ 0, \text{ else} \end{cases}. \tag{9}$$

In II we consider the transformation

$$\widetilde{\mathbf{p}}_{bp}(t) = \mathbf{p}_{bp}(t) - \min \mathbf{p}_{bp}(t) + 1/N \tag{10}$$

which yields a vector $\widetilde{\mathbf{p}}_{bp}(t)$ in which the smallest value is $1/N$. Furthermore, consider the normalized vector

$$\mathbf{v}(t) = \frac{\widetilde{\mathbf{p}}_{bp}(t)}{\text{sum}\ \widetilde{\mathbf{p}}_{bp}(t)} \tag{11}$$

and also the uniform probability vector

$$\overline{\mathbf{v}}(t) = \left[\frac{1}{N}\ \frac{1}{N}\ \cdots\ \frac{1}{N}\right]^T_{1\times N} \tag{12}$$

where N is the size of $\mathbf{p}_{bp}(t)$ (N=256 in this case). Note that the vectors found, $\mathbf{v}(t)$ and $\overline{\mathbf{v}}(t)$, can now be considered to be probability vectors. Clearly, if $\mathbf{v}(t)$ and $\overline{\mathbf{v}}(t)$ are similar there is no clear peak in $\mathbf{v}(t)$ and the block are not to be considered as a whistle block. The obvious question is now how to create a similarity measure. One way is to use the *Jensen difference* [2, 12]. The Jensen difference is based on the *Shannon entropy* (the block index t is dropped for simplicity)

$$H(\mathbf{v}) = -\sum_{i=1}^{N} v_i \log(v_i) \tag{13}$$

and is defined as

$$J(\mathbf{v}, \overline{\mathbf{v}}) = H\left(\frac{\mathbf{v}+\overline{\mathbf{v}}}{2}\right) - \frac{1}{2}\left(H(\mathbf{v}) + H(\overline{\mathbf{v}})\right). \tag{14}$$

The Jensen difference is always nonnegative and becomes zero only if $\mathbf{v} = \overline{\mathbf{v}}$ [12]. To simplify notation is $J(\mathbf{v}(t), \overline{\mathbf{v}}(t))$ denoted by $J(t)$. The decision for accepting a block as a whistle block is made by setting a threshold on the Jensen difference, that is

$$b(t) = \begin{cases} 1, \text{ if } J(t) > \theta_J \\ 0, \text{ else} \end{cases}. \tag{15}$$

To summarize I and II, and thereby checking for the statement in III, the detection function $c(t)$ is found as

$$c(t) = a(t)b(t) \tag{16}$$

which implies that both $a(t)$ and $b(t)$ should be one for the whistle decision to be valid. Further, the decision function $c(t)$ will be smoothed over time to make IV valid. This smoothing over the last D blocks is performed by

$$d(t) \begin{cases} 1, \text{ if } \left(\sum_{\tau=t-D+1}^{t} c(\tau)\right) > D/2 \\ 0, \text{ else} \end{cases} \tag{17}$$

where $d(t)$ constitutes the final whistle detection function.

6 Experiments

The system described in section 4 is implemented and runs in real-time on a Pentium 2.13GHz computer. The parameters $\theta_\gamma = 25$, $\theta_J = 0.45$ and $D = 50$ are experimentally chosen. During the experimentations, different background noises have been tested, including music, white noise, babble and car noise, see Fig. 6. In order to compare the whistle detector the actual whistle location, denoted $d_o(t)$, is included in the figure. The detection is typically delayed with a few blocks due to the smoothing according to Eq. (17). Also different microphones and sound-cards were used to analyze the performance. The parameters chosen are found to be robust to changes in sensor and platform, this mainly due to the use of the SMQT. Some typical false detections occur in the presence of music with long duration tones in the frequency band 500-5000 Hz (which are very similar to whistle). Some single frequency screaming have been found to cause false detection in some cases.

Hardware consisting of a simple microcontroller, high voltage relay and numerous analog and digital components was constructed. The hardware acts as an electronic switch controlling the electricity flow in the cable. The switch is controlled by the computer through COM serial port interface. A video clip www.asb.tek.bth.se/staff/jsb/whistle/whistle.html created by the authors demonstrates a live scenario where a lamp is turned on/off by the whistle in the room. Simultaneously different types of sounds are played in the room acting as a noise source. This particular hardware can also be used in other applications where various devices need a power supply.

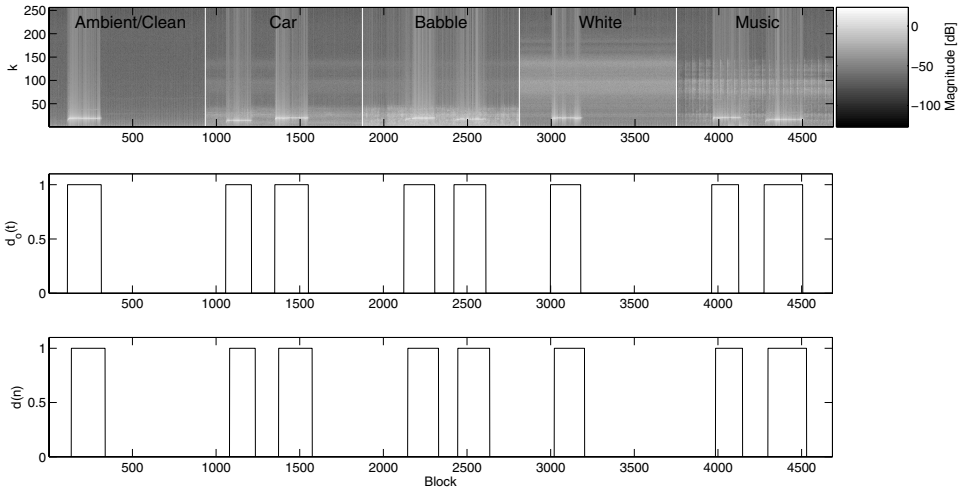

Figure 6. Human whistle signal in various noise situations.

7 Conclusion

Human whistling was investigated from a database with collected whistle sounds. The typical frequency range for human whistle was found to be 500-5000 Hz. Given this knowledge a feature extraction technique was proposed. The feature vectors was further analyzed to achieve detection and frequency estimation of human whistling. The final system runs at real-time and was capable of detecting human whistle during various noise situations.

References

[1] M. Böhlen and J. T. Rinker. Unexpected, Unremarkable, and Ambivalent OR How The Universal Whistling Machine Activates Language Remainders. In *Computational Semiotics for Games and New Media, COSIGN2004*, 2004.

[2] J. Burbea and C. Rao. On the Convexity of Some Divergence Measures Based on Entropy Functions. *IEEE Trans. Information Theory*, IT-28(3):489–495, 1982.

[3] Y. Chan, Q. Ma, H. So, and R. Inkol. Evaluation of various fft methods for single tone detection and frequency estimation. In *IEEE 1997 Canadian Conference on*, volume 1, pages 211–214, May 1997.

[4] J. Dubnowski, J. French, and L. Rabiner. Tone detection for automatic control of audio tape drives. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 24, pages 212–215, June 1976.

[5] P. J. G. and M. D. G. *Digital Signal Processing*. Prentice-Hall, 1996. ISBN 0-13-394338-9.

[6] H. M. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley & Sons Inc., 1996. ISBN 0-471-59431-8.

[7] K. Kanagisawa, A. Ohya, and S. Yuta. An operator interface for an autonomous mobile robot using whistle sound and a source direction detection system. In *Proceedings of the 1995 IEEE IECON 21st International Conference on Industrial Electronics, Control, and Instrumentation*, volume 2, pages 1118–1123, November 1995.

[8] S. Lefevre, B. Maillard, and N. Vincent. 3 classes segmentation for analysis of football audio sequences. In *14th International Conference on Digital Signal Processing*, volume 2, pages 975–978, July 2002.

[9] M. Nilsson, M. Dahl, and I. Claesson. The successive mean quantization transform. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 429–432, March 2005.

[10] M. Nilsson, M. Dahl, and I. Claesson. Gray-scale image enhancement using the SMQT. Accepted and presented at IEEE International Conference on Image Processing (ICIP), Genova 2005.

[11] P. Tyack, W. Williams, and G. Cunningham. Time-frequency fine structure of dolphin whistles. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 17–20, October 1992.

[12] R. Vergin and D. O'Shaughnessy. On the Use of Some Divergence Measures in Speaker Recognition. In *Proceedings of ICASSP*, pages 309–312, 1999.