A TONE-FIT FEATURE REPRESENTATION FOR SCENE CLASSIFICATION

Johannes D. Krijnders, Gineke A. ten Holt

INCAS^3
Dr. Nassaulaan 9
9401 HJ Assen
The Netherlands
{gineketenholt, dirkjankrijnders}@incas3.eu

ABSTRACT

We present an algorithm that classifies environmental sound recordings using a feature representation based on the human hearing. Specifically, we use a mathematical model of the human cochlea to transform a sound (wav) clip into a time-frequency representation called a *cochleogram*. From the cochleogram, we calculate the tone-fit of each time-frequency region by calculating the fit of the region to a pure tone. This gives us a representation of the 'tone-likeness' of the sound at various moments and frequencies. Finally, to arrive at a summarized representation for the entire clip, we calculate 20 statistic components over the tone-fit matrix. The resulting 20-dimensional feature representation is then classified using a support vector machine. The accuracy of the resulting method is 0.53 ($SE = 0.06$). Similar results are obtained by using MFCC features and voting by frame ($0.60, SE = 0.04$). Future directions include separately identifying sound events and representing scenes in terms of component events.

Index Terms— Cochlea model, environmental sound recognition, auditory scene classification

1. INTRODUCTION

The challenge is to classify real-world auditory scenes into ten predefined classes. The dataset consists of 30-second recordings, ten examples per class [1]. Our approach is based on a transmission-line model of human cochlea [2] and a representation of the tonalness of the resulting energy representation. The cochlea model presents a time-frequency transformation that allows for a optimal resolution and noise-robustness [3]. The idea behind the tone-fit representation is that sound that is purposefully produced is often tonal to ensure a good local signal-to-noise ratio [4].

2. COCHLEOGRAM

The features for the classification are calculated on a time-frequency representation based on a model of the human cochlea. The model taken is the transmission-line model of Duifhuis et al. [2]. The model discretizes the basilar membrane in 200 equal segments. Each segment is modeled as a mass-spring-damper system(eqn. 1), which is solved by gauss elimination and a fourth order runge-kutta time integration [5].

$$\frac{\delta^2\left(m\frac{\delta^2 y}{\delta t^2}+d(x)\frac{\delta y}{\delta t}+s(x)y\right)}{\delta x^2}-\frac{2\rho}{h}\frac{\delta^2 y}{\delta t^2}=0 \qquad (1)$$

The characteristic frequencies of the segments as described by Greenwood[6]:

$$f_c(x) \quad = \quad 2\left(A10^{-ax}-f_0\right) \qquad (2)$$

where $A = 17.927kHz$, $a = 60/m$, $f_0 = 145.4Hz$ and $0 < x < 35mm$. This relation determines the damping component $d(x)$ for a fixed $s(x)$. For parameter values see [5]).

To calculate an energy representation the output is squared and leaky integrated with a channel dependent time-constant $\tau_c = \max(5, 2/f_c)$ ms. The integrated output is down-sampled to 200 Hz and compressed logarithmically to express the energy in dB. The resulting representation is a spectrogram-like representation, termed a cochleogram, with 5 ms frames.

3. TONE-FIT ALGORITHM

After converting the time signal to the cochleogram domain, the tonalness of each time-frequency point is calculated. We apply channel dependent matched filters that respond to ideal tones and pulses. The derivation of these filters is depicted in Fig. 1. For each channel an ideal sinusoid is generated and processed using the cochlea model (Fig. 1a). Subsequently the width of the response in frequency at a threshold under the energy maximum is calculated (Fig. 1c). This threshold is set to twice the standard deviation of the log energy of white noise in a channel. This standard deviation is independent of the power spectral density of the noise in the logarithmic energy domain. The width of the response is the filter parameter for the tone-fit (TF). Application (Fig. 2) of the filters is the complementary process, the energies at the widths below and above the timefrequency point are averaged. The difference between the energy at the point for that channel and the average forms the filter output. The application of the filter to the cochleogram results in the tone-fit representation like in figure 3.

INCAS^3 is co-financed by the European Union (European Fund for Regional Development), the Dutch Ministry of Economic affairs (Peaks in the Delta), the Province of Drenthe and the Municipality of Assen.

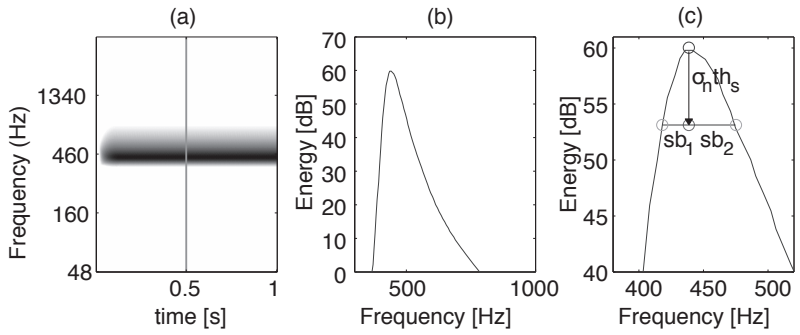

Figure 1: Calculation of TF filters. (a) Cochleogram of ideal tone. (b) Cross-section at $t = 0.5$s. (c) Detail of (b) with filter parameters.

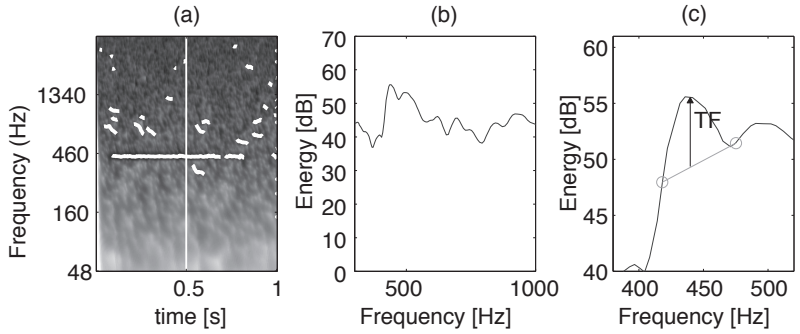

Figure 2: Application of TF filters. (a) Cochleogram of ideal tone in zero dB local SNR white noise. (b) Cross-section at $t = 0.5$s. (c) Detail of (b) with filter parameters.

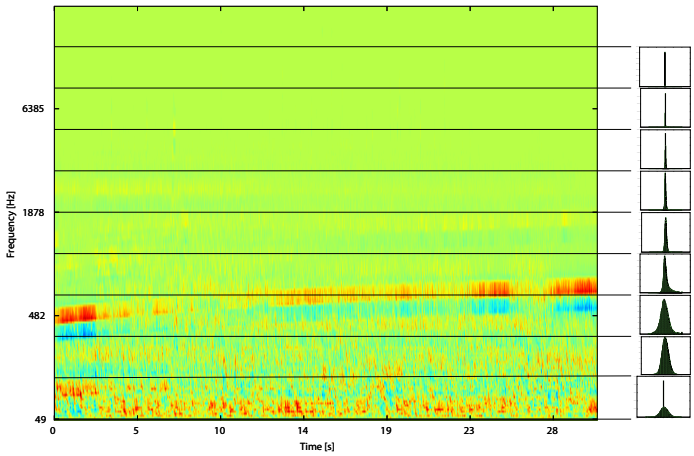

Figure 3: Example tone-fit matrix, the horizontal lines indicate the parts over which the features are extracted. The 5% and 95% points in the histograms on the right are the features used.

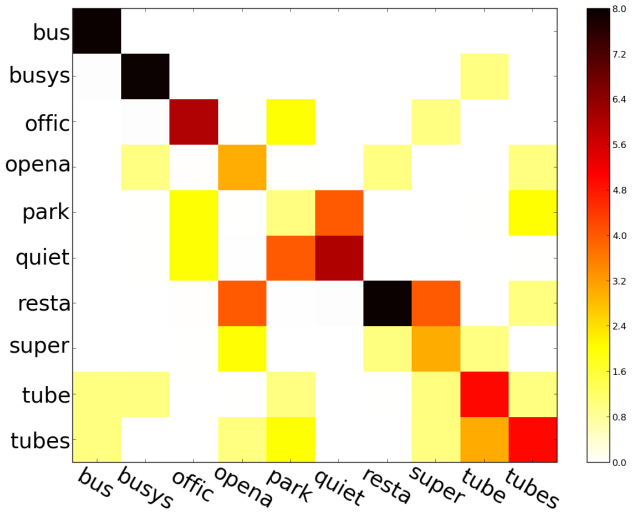

Figure 4: Confusion matrix for 5-fold cross-validation on tone-fit features plus spectral slope, using a support vector machine with a radial basis function kernel. Rows are the predicted labels, columns are the actual labels.

4. FEATURE REPRESENTATION

From the tone-fit representation 20 features are calculated. The segments are divided in ten equal parts and for each part a histogram of all tone-fit values is made, see figure 3. The tone-fit value that is exceeded 5 percent of time and the value that is exceeded 95 percent of the time form 20 features. Due to edge effects, the features for the segments corresponding to the highest frequencies are unreliable, and are discarded. The overall mean and standard deviation form features 19 and 20. Finally, two features representing the spectral slope (slope and offset) are added, giving a total of 22 features.

5. CLASSIFICATION

With the sound samples thus represented as 22-dimensional vectors, a support vector machine (SVM) is trained on them [7]. On the development dataset, the accuracy for 5-fold cross-validation was $0.53(SE = 0.06)$. Figure 4 shows the confusion matrix for SVM classification using a radial basis function kernel. Classes 'bus', 'busystreet' and 'restaurant' perform rather well, perhaps due to tonal qualities (voices in the restaurant?) not found in other classes. Mutual confusion exists between 'park' and 'quietstreet' and between 'tube' and 'tubestation'; not surprising since the two will have general features in common (open air quietness, sounds of trains and stations).

6. DISCUSSION AND FUTURE DIRECTIONS

Apart from tone-fit and SVM, we have investigated a variety of feature representations and classification methods, including MFCCs and voting by frame, sound level representations, and automatic clustering of foreground energy features. Of these approaches, tone-fit and MFCCs performed best, but still only achieved around 55% accuracy. We find that the MFCC+GMM as used in the base-line system [8] makes different errors then the presented system with the tone-fit, combining the the two may improve the results.

What all these methods have in common is that they attempt, in one way or another, to arrive at a generalized description of the auditory scene. It would appear that a description general enough to capture the variety of sounds that can occur in a certain class quickly becomes too general to distinguish between similar classes.

Interesting future directions for research would therefore be: combining general classifiers, so that the weaknesses of one representation can be mitigated by others; or a representation that instead of generalizing over the entire recording, would isolate and label elements within the recording (either events or sub-event elements), for example [9]. In the latter case, a recording could then be represented in terms of presence and importance of elements. A drawback of this approach is that it requires the annotation of the sources present in the scene for the training data, which is very time consuming.

7. REFERENCES

[1] http://www.elec.qmul.ac.uk/digitalmusic/sceneseventschallenge/.

[2] H. Duifhuis, H. W. Hoogstraten, S. M. Van Netten, R. J. Diependaal, and W. Bialek, *Peripheral Auditory Mechanisms*, ser. NATO ASI SERIES, J. P. Wilson and D. T. Kemp, Eds. Springer Verlag, 1986.

[3] P. W. J. van Hengel, "A Comparison of Spectro-temporal Representations of Audio Signals," in *IAI/DAGA 2013*, Merano, Mar. 2013.

[4] J. D. Krijnders, "Signal-driven sound processing for uncontrolled environments," Ph.D. dissertation, Rijksuniversiteit Groningen, Groningen, Oct. 2010.

[5] R. J. Diependaal, "Numerical methods for solving one-dimensional cochlear models in the time domain," *The Journal of the Acoustical Society of America*, vol. 82, no. 5, p. 1655, 1987.

[6] D. D. Greenwood, "Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane," *The Journal of the Acoustical Society of America*, Jan. 1961.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.

[8] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *Proceedings of EUSIPCO 2013*, 2013.

[9] M. E. Niessen, J. D. Krijnders, and T. C. Andringa, "Understanding a soundscape through its components," in *Euronoise*, Edinbourgh, Oct. 2009.