# TEMPORAL CODING OF LOCAL SPECTROGRAM FEATURES FOR ROBUST SOUND RECOGNITION

*Jonathan Dennis, Yu Qiang, Tang Huajin, Tran Huy Dat, Li Haizhou*

Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632

**ABSTRACT**

There is much evidence to suggest that the human auditory system uses localised time-frequency information for the robust recognition of sounds. Despite this, conventional systems typically rely on features extracted from short windowed frames over time, covering the whole frequency spectrum. Such approaches are not inherently robust to noise, as each frame will contain a mixture of the spectral information from noise and signal. Here, we propose a novel approach based on the temporal coding of Local Spectrogram Features (LSFs), which generate spikes that are used to train a Spiking Neural Network (SNN) with temporal learning. LSFs represent robust location information in the spectrogram surrounding keypoints, which are detected in a signal-driven manner such that the effect of noise on the temporal coding is reduced. Our experiments demonstrate the robust performance of our approach across a variety of noise conditions, such that it is able to outperform the conventional frame-based baseline methods.

***Index Terms***— Sound recognition, neural coding, local features

## 1. INTRODUCTION

The task of recognising sounds in noisy and unstructured environments is a major challenge faced by the audio processing field [1]. Recently there has been renewed interest on the topic of "machine hearing" [2], where the aim is to be able to achieve human-like recognition performance across a wide range of sounds and signal-to-noise ratios (SNRs). Conventional approaches are typically "frame-based", which models the acoustic signal as a series of fixed dimension features extracted from each time frame of the continuous audio. The most commonly used features are Mel-Frequency Cepstral Coefficients (MFCCs) [3], which are often modelled using Gaussian Mixture Models (GMMs), with the temporal information captured using Hidden Markov Models (HMMs).

However, there are two significant drawbacks of such frame-based approaches. Firstly, sounds display a wide variety of spectral characteristics, with many examples that contain relatively sparse frequency spectrums with most energy contained in a few frequency bands. When noise is present in the signal, the noise will mask the spectral information of the sound in certain regions of the spectrogram, through the LogMax principle [4], with each frame containing a mixture of information from both the noise and sound. This can cause the frame to become arbitrarily far from the GMM trained on only clean instances of the same sound, and hence significantly reduce the performance. Secondly, typically HMMs do not explicitly model the temporal coding of the underlying frames. Instead, they rely on a first order model, based on the transition probability from the previous frame. However, sounds have a much more diverse temporal dependency, such that a more complete modelling of the temporal information should improve the performance.

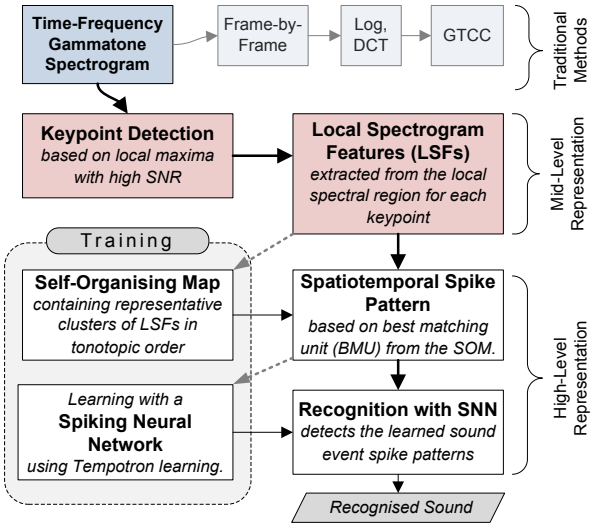



**Fig. 1**. Overview of the proposed LSF-SNN recognition system, compared to traditional auditory methods such as the Gammatone Cepstral Coefficients (GTCC).

In this paper, we propose the novel LSF-SNN approach for sound recognition, based on the temporal coding of Local Spectrogram Features (LSFs), using a spiking neural network (SNN) with temporal learning for recognition. In a previous work [5], we utilised the idea of LSFs for overlapping sound recognition using the Generalised Hough Transform for sound source separation. Here, our purpose is to develop the LSF into a biologically inspired system for robust sound recognition, which is a departure from our previous work. The idea here is that even when noise is present, we can extract a reliable LSF from the spectral region surrounding "keypoints" which are detected on the sparse, high SNR peaks in the two-dimensional sound spectrogram. These keypoints are detected in a signal-driven manner that is independent of the sound class or noise condition, such that a temporal coding based on these keypoints will be robust when noise is present. To generate the temporal code, our solution is to model characteristic representations of the LSF information in an unsupervised manner using Self Organising Maps (SOM) during training [6], which can be seen as an tonotopic (frequency ordered) mid-level representation of the sound information. The Best Matching Unit (BMU) of the SOM then generates a spike at the time that the keypoint occurred to form a spatiotemporal spike pattern, which can be learnt in an SNN for recognition of the underlying sound. An overview of this system is found in Fig. 1.

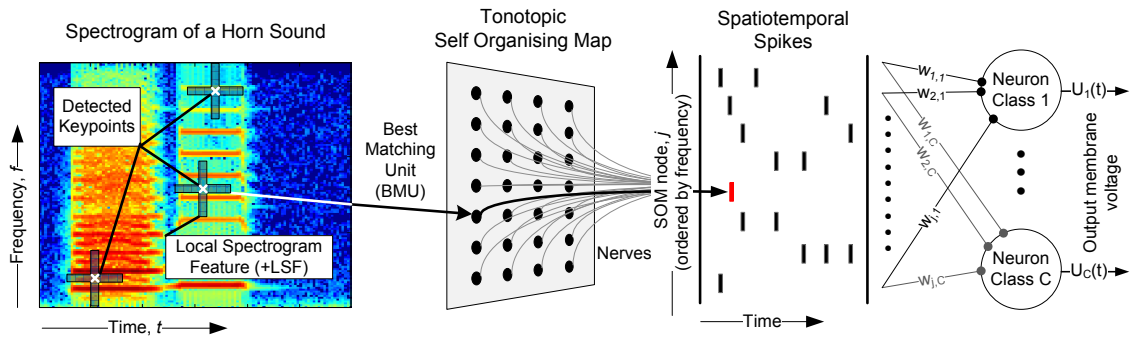



**Fig. 2**. The proposed LSF-SNN system. First we detect keypoints and extract LSFs, followed by the SOM mapping to produce the output spatiotemporal spike patterns. The weights, $w_{i,c}$ are then learnt by a SNN using Tempotron learning for recognition.

Our motivation is based on biological evidence, that suggests the human auditory does not process acoustic information in such a frame-based way. Rather, it has been suggested that the auditory system processes audio in local frequency bands, such that it can recognise noise corrupted audio based on local time-frequency regions with high SNR [7, 8]. This forms the idea for using the LSF as our mid-level representation, as it captures only the local spectral information, which is more robust to noise than a frame-level representation. In addition, it is known that neurons communicate with each other by means of short spikes, thereby representing external stimuli in the brain in a form of spatiotemporal spike patterns [9]. In some cases neurons are commonly assumed to represent information by mean firing rate, however neurons in both the visual [10] and auditory [11] pathways are observed to precisely respond to the stimulus on a millisecond timescale [12]. These results support the hypothesis of temporal coding, where precise timings of spikes are taken into account for conveying information.

However, while many previous works have considered a biologically inspired auditory front-end, such as the Gammatone filterbank [13], most have simply fallen back on the traditional pattern recognition approaches such as the Gammatone Cepstral Coefficients (GTCC) in [14], rather than creating a more complete biological system as we have focussed on here. Others have also utilised the properties of SOMs for speech recognition, [15], for example by utilising the BMU-trajectory, and have used SNNs for biologically inspired sound recognition [16]. However, these approaches often do not consider the recurring problem of robustness to noise, and have also not been combined in such a way as the novel system presented here. Other works have also utilised keypoints in the spectrogram, particularly for music identification such with the Shazam system [17], which hashes together pairs of keypoints as the basis for recognition. However, the approach does not extract a local feature, and in addition it has been shown in [18] that such a system is not well suited to general sounds, due to the precise nature of the features that may not be reliably repeated. Therefore, in this paper we compare our approach to a conventional frame-based MFCC-HMM system for noise-robust recognition using missing features. Our experiments demonstrate the robustness of our approach compared to this well-performing baseline technique.

The rest of this paper is as follows: Section 2 details our proposed LSF-SNN approach for learning spatiotemporal patterns based on LSFs. Section 3 then describes the experiments used to validate our approach, before Section 4 concludes the work.

## 2. PROPOSED LSF-SNN SOUND RECOGNITION SYSTEM

### 2.1. Signal-driven Local Spectrogram Feature Extraction

We start be representing the audio signal as a log-power Gammatone spectrogram, $S(f,t)$, where $f = 1 \dots F$ is the centre frequency of the Gammatone filter and $t$ is the time frame after down-sampling the spectrogram into 16ms frames with 50% overlap. We use audio clips with a 16 kHz sampling frequency, and $F = 50$ filters spaced equally on the Equivalent Rectangular Bandwidth (ERB) scale.

The basis for both our keypoint detection and LSF extraction is the plus-shaped local spectrogram region, which we show on the left of Fig. 2. The idea is to capture the local spectral and temporal shape, such that it gives a "glimpse" [8] of the local spectral information in two dimensions. We found previously [5] that this is more suitable than including the full 2D region, which may contain a significant amount of non-stationary noise. The plus-shaped region is composed of the local horizontal and vertical spectrum, as follows:

$$\begin{aligned} Q_f(y) &= S(f \pm d, t), \quad d = [1, 2, \dots, D] \\ Q_t(y) &= S(f, t \pm d) \end{aligned} \tag{1}$$

where $Q_f, Q_t$ are the local spectral and temporal vectors respectively, $y = [1, 2, \dots, 2D]$ is the vector index, and $D = 6$ is the half-width of the local region, which we found was small enough to extract the important local peaks, but large enough to provide a feature for clustering and classification.

Keypoints are then detected at locations that are local maxima across either frequency or time, which ensures they can be detected on both short impulsive sounds that appear as vertical lines in the spectrogram, as well as on harmonic sounds that appear as horizontal lines. A keypoint is then detected if:

$$S(f,t) > \begin{cases} Q_f(y), \ or \\ Q_t(y). \end{cases} \quad \forall y = [1, 2, \dots, 2D] \tag{2}$$

The output is a set of keypoint information, $K_i$, as follows:

$$K_i = \{f_i, t_i, s_i, L_i, M_i\} \tag{3}$$

where $f_i, t_i$ are the time-frequency coordinates, $s_i = S(f_i, t_i)$ is the spectral power, and $L_i, M_i$ are the LSF and missing feature mask respectively, which are detailed below.

The LSF is formed by concatenating $Q_{f,i}$ and $Q_{t,i}$, such that it represents the plus-shaped local region, hence we call this the +LSF.

This is normalised by the spectral power, $s_i$, such that the +LSF, $L_i$, is written as follows:

$$L_i = \frac{\{Q_{f,i}, Q_{t,i}\}}{s_i} \tag{4}$$

The +LSF hence characterises the local spectral shape, independent of the relative magnitude of the sound, which leads to a more compact set of LSF representations, as similar patterns of different magnitude can be clustered together.

Next, we generate a local missing feature mask [19] for each +LSF to prevent noise from affecting the matching of LSFs between training and testing. To estimate this we assume that the noise will be stationary in the local region across either the spectral, $Q_{f,i}$, or temporal, $Q_{t,i}$, dimension. The local noise estimate, $\mu_i$, is then simply the minimum of the two means, as follows:

$$\mu_i = \min\left[mean(Q_f),\ mean(Q_t)\right]. \tag{5}$$

The missing feature mask, $M_i$, is then formed as follows:

$$M_i(z) = \begin{cases} 1, & \text{for } L_i(z) > \mu_i \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $z = 1 \ldots 4D$ is a variable representing the +LSF dimensions.

Finally, we define a local sparsity measure, $\delta_i = s_i - \mu_i$, to reject less significant keypoints that are more likely to belong to noise rather than signal. Here, $\delta_i$ controls the minimum size of the peak for it to be significant, which we set to $\delta_i > 5$dB based on preliminary experiments. This reduces the computation required during recognition, and reduces the chance of a false match occurring due to fluctuations in the noise.

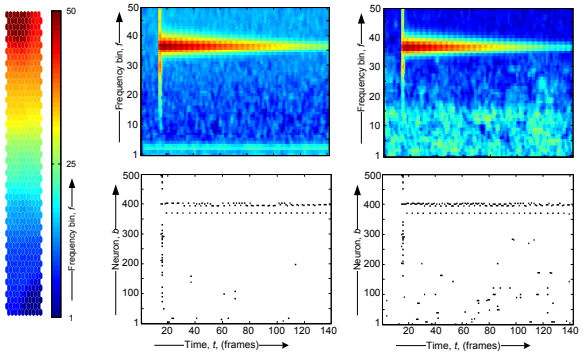
### 2.2. Temporal Coding of LSFs using Self Organising Maps

During training, we perform LSF clustering on the extracted keypoint information, $K_i$, to learn clusters of similar input features. In our previous work [5], we used k-means clustering. However, here we choose to use Self-Organising Maps (SOM) [6], which is an unsupervised, biologically plausible competitive learning process. We do this as the trained SOM contains ordered clusters, due to its neighbourhood learning rule, which we use to produce the tonotopic ordering of the neurons that is seen in the human auditory system [20]. We do this by appending the keypoint frequency information, $f_i$, to the feature to give $L'_i = \{L_i, f_i\}$, which enables the SOM to produce a tonotopic topology. An example of the learnt tonotopic map is shown in Fig. 3a, with the frequencies arranged in ascending order.

We use the SOM Toolbox for Matlab [21] for the learning process, with the SOM set to have a rectangular map of size $N = 50 \times 10$ units, and a Gaussian neighbourhood function. To calculate the Best Matching Unit (BMU), $b_j$, of the SOM, $X$, we use the squared Euclidean distance measure between the input vector, $L'_i$, and each unit's pattern, $x_n$. This is weighted by the masking function, $M_i$, extracted from each keypoint, and can be written as:

$$b_j = \min_n \left[(x_n - L'_i)^T M_i^{-1} (x_n - L'_i)\right], \quad \forall n \in X \tag{7}$$

The BMU, $b_j$, represents the occurrence of the +LSF, $L'_i$, at time $t_i$ specified by the position of the keypoint, $K_i$. Over all the keypoints detected in the spectrogram, this generates a spatiotemporal spike pattern, $P(b,t)$, where each keypoint-BMU match generates a spike by setting $P(b_j, t_i) = 1$ The output of the SOMs forms the spatiotemporal spike pattern as shown in Fig. 2. The pseudo code for the temporal coding can be written as shown in Algorithm 1.




(a) SOM Tonotopic mapping.

(b) Bell sound in clean conditions *(right)* and 10dB noise *(left)* with the corresponding spatiotemporal spike pattern shown below to demonstrate the noise robustness.

**Fig. 3**. Examples of the SOM mapping and spatiotemporal patterns.

**Algorithm 1** Temporal Coding of LSFs

Generate a Log-power Gammatone spectrogram: $S(f,t)$
Extract set of keypoints: $K_i = \{f_i, t_i, s_i, L_i, M_i\}$.
Reset the spatiotemporal spike pattern, $P(b,t) \Leftarrow 0$
**for all** $K_i$ **do**
  Find BMU, $b_i$, for $[L'_i, M_i]$, in SOM $X$
  Set a spike in the spatiotemporal pattern: $P(b_j, t_i) \leftarrow 1$
**end for**

Examples spike trains are shown in Fig. 3b for a bell sound in both clean and 10dB noise. It can be seen that each sound generates a distinctive pattern that represents the information in the spectrogram through the time-frequency occurrences of +LSF patterns learnt by the tonotopic SOM. It can also be seen that the clean and noisy patterns are very similar. While there are some random spikes detected due to the noise, the important bell information is still represented, therefore the spatiotemporal coding is robust.

### 2.3. Temporal Learning Rule

Here, we describe the learning rule we used for processing the spatiotemporal patterns. Temporal learning aims at dealing with information encoded by precise timing spikes. As proposed in [22], the tempotron rule is efficient for classifying a great number of spatiotemporal patterns. This biologically plausible rule has previously been successfully applied in a simple task of time-warp-invariant word discrimination [23], hence we adopt this rule here.

According to the tempotron rule, the synaptic plasticity is governed by the temporal contiguity of a presynaptic spike and a postsynaptic depolarisation, and a supervisory signal. This rule modifies the synaptic weights such that the trained neuron will emit one spike when it is presented with a pattern corresponding to one category ($C^+$) and no spike presented with a pattern corresponding to another category ($C^-$). The subthreshold membrane voltage $U(t)$ of the neuron is a weighted sum of postsynaptic potentials (PSPs) from all incoming spikes:

$$U(t) = \sum_j w_j \sum_{t_i < t} S(t - t_i) + U_{rest} \quad \forall\, t \in [0, T] \tag{8}$$

where $w_j$ and $t_i$ are the synaptic efficacy and the fired time of the

$j^{th}$ afferent, and $T$ is the duration of the pattern. $U_{rest}$ is the rest potential of the neuron. $S$ denotes the normalised PSP kernel:

$$S(t-t_i) = S_0(\exp(\frac{-(t-t_i)}{\tau_m}) - \exp(\frac{-(t-t_i)}{\tau_s})) \quad (9)$$

where $\tau_m$ and $\tau_s$ are decay time constants. $S_0$ normalises PSP so that the maximum value of the kernel is 1.

Without incoming spikes, the neuron's potential is at rest. Each incoming spike will trigger the change of neuron's potential, and the maximum change by this spike depends on its synaptic efficacy. When $U(t)$ crosses the threshold, the neuron emits a spike, after which the potential gradually decreases to the rest value by shunting all the following input spikes.

For classifying two categories, the neuron will modify its synaptic weights whenever there is an error response. The learning rule used depends on the firing of the neuron, which either (A) fails to fire, or (B) erroneously fires:

$$\Delta w_j = \begin{cases} \lambda \sum_{t_i<t_{max}} S(t_{max}-t_i), & \text{if A;} \\ -\lambda \sum_{t_i<t_{max}} S(t_{max}-t_i), & \text{if B;} \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

where $t_{max}$ denotes the time at which the neuron reaches its maximum potential value, and $\lambda > 0$ is the learning rate. In this learning rule, long term potentiation (LTP) is activated if the neuron failed to spike on a $C^+$ pattern and long term depression (LTD) is activated if the neuron erroneously fired a spike on a $C^-$ pattern.

In the multi-class case, we train one neuron corresponding to each category. The testing pattern is classified to the category that associates with the most strongly activated neuron.

## 3. EXPERIMENTS

In this section we carry out experiments to show the performance of our proposed LSF-SNN system on a sound recognition task.

**Sound Database:** We select the following ten sound classes from the Real Word Computing Partnership Sound Scene Database [24]: bells5, bottle1, buzzer, cymbals, horn, kara, metal15, phone4, ring and whistle1. The sound files have a high SNR, and each contains an isolated sound, with a small amount of silence before and after. For each of the 10 classes, 20 files are randomly selected for training and another 20 for testing, giving 400 clips in total. The average performance is then reported across 5 runs of the experiment.

**Experimental Methods:** We compare our method against two baseline frame-based HMM systems, including one that uses missing features to cope with the different noise conditions [19]. The idea is to estimate a reliability mask for the spectrogram and modify the classifier to deal with these missing elements. Here we use bounded marginalisation, using the implementation provided in the CASA Toolkit [25]. This aim is to provide a comparison between the approach in this paper utilising a local missing feature mask, as estimated for each LSF feature in (6), and a conventional missing feature approach that is applied to frame-based features.

The following methods are therefore evaluated:

1. Proposed LSF-SNN method, with an SOM comprised of $N = 50 \times 10 = 500$ neurons, which we found gave a good trade-off between accuracy and performance.
2. Baseline MFCC-HMM, with 5 states and 5 Gaussian mixtures per state. The frame-based MFCCs have 36-dimensions, with 12 cepstral coefficients without the zeroth component, plus their deltas and accelerations.
3. MF-HMM, using a 36-dimension Mel-frequency spectral coefficient feature, with the same HMM configuration as above, but with the decoder supporting missing feature (MF) marginalisation. We estimate the mask using the local SNR method, as given in [19].

| *Method:* | LSF-SNN | MFCC-HMM | MF-HMM |
|---|---|---|---|
| Clean | 98.5% | **99.0%** | 95.7% |
| 20dB | **98.0%** | 62.1% | 94.2% |
| 10dB | **95.3%** | 34.4% | 84.7% |
| 0dB | **90.2%** | 21.8% | 69.5% |
| -5dB | **84.6%** | 19.5% | 53.8% |
| Average | **93.3%** | 47.3% | 79.6% |

**Table 1**. Results for the proposed LSF-SNN method against two baseline HMM systems, including a missing feature (MF) method.

For this experiment, the classification accuracy is investigated in mismatched conditions, using only clean samples for training. The average performance for each method is reported in clean and at 20, 10, 0 and -5 dB signal-to-noise ratio (SNR) for the "Speech Babble" noise environment, taken from the NOISEX'92 database [26].

**Results:** The experimental results are presented in Table. 1. It can be seen that the proposed LSF-SNN method performs well for each of the noise conditions, achieving an average accuracy of $93.3\%$ down to -5dB. It can also maintain an accuracy of over $90\%$ in the challenging 0dB SNR condition. This outperforms the two baseline methods in all but the cleanest condition. Although even here, we still achieve a very good accuracy of $98.5\%$, which is only $0.5\%$ lower than the baseline MFCC-HMM, which was trained in matched conditions.

The results also show that the simple frame-based MFCC-HMM is not robust to noise, as it achieves an average performance of only $47.3\%$. This is because there is no compensation for the mismatch between clean training samples and the noisy testing conditions. The MF-HMM missing feature method improves considerably on this result, achieving an average performance of $79.6\%$ over the different SNR conditions. However, in clean conditions the method did not perform as well as either the MFCC-HMM or our proposed LSF-SNN methods, and in noisy conditions the performance is limited by the accuracy of the missing feature mask, which is difficult to estimate reliably over the whole spectrogram.

Overall, our proposed LSF-SNN method performed well in both clean and noisy conditions, demonstrating a robust performance across a wide range of challenging noise conditions.

## 4. CONCLUSION

In this paper, we proposed the LSF-SNN system for robust sound recognition, which is a novel approach based on the temporal coding of local information in the spectrogram. We made use of local spectrogram features, which we found can characterise the information in the spectrogram well and, through a local missing feature mask, is also robust in the presence of background noise. These were then clustered using a tonotopic SOM to produce temporal spike patterns, which represent the time-frequency occurrences of LSFs in the spectrogram, that could then be learnt using an SNN using the Tempotron learning rule. Our experiments showed that our LSF-SNN system could outperform the best-performing baseline approach in all but the cleanest of conditions, which underlies the effectiveness of the approach.

## 5. REFERENCES


[1] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[2] R.F. Lyon, "Machine hearing: An emerging field," *Signal Processing Magazine, IEEE*, vol. 27, no. 5, pp. 131–139, 2010.

[3] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[4] S.T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[5] J. Dennis, H.D. Tran, and E.S. Chng, "Overlapping sound event recognition using local spectrogram features with the generalised hough transform," in *Thirteenth Annual INTERSPEECH Conference*, September 2012.

[6] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[7] J.B. Allen, "How do humans process and recognize speech?," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 567–577, 1994.

[8] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, pp. 1562, 2006.

[9] Fred Rieke, David Warland, Rob, and William Bialek, *Spikes: Exploring the Neural Code*, MIT Press, Cambridge, MA, 1st edition, 1997.

[10] Tim Gollisch and Markus Meister, "Rapid neural coding in the retina with relative spike latencies.," *Science*, vol. 319, no. 5866, pp. 1108–1111, 2008.

[11] R C deCharms and M M Merzenich, "Primary cortical representation of sounds by the coordination of action-potential timing.," *Nature*, vol. 381, no. 6583, pp. 610–613, 1996.

[12] Daniel A. Butts, Chong Weng, Jianzhong Jin, Chun-I Yeh, Nicholas A. Lesica, Jose-Manuel Alonso, and Garrett B. Stanley, "Temporal precision in the neural code and the timescales of natural vision," *Nature*, vol. 449, no. 7158, pp. 92–95, Sept. 2007.

[13] RD Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *APU report*, vol. 2341, 1988.

[14] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *Multimedia, IEEE Transactions on*, vol. PP, no. 99, pp. 1, 2012.

[15] P. Somervuo and T. Kohonen, "Self-organizing maps and learning vector quantization for feature sequences," *Neural Processing Letters*, vol. 10, no. 2, pp. 151–159, 1999.

[16] J.J. Hopfield and C.D. Brody, "What is a moment?cortical sensory integration over a brief interval," *Proceedings of the National Academy of Sciences*, vol. 97, no. 25, pp. 13919, 2000.

[17] A. Wang, "An industrial strength audio search algorithm," in *International Conference on Music Information Retrieval (ISMIR)*, 2003, vol. 2.

[18] J.P. Ogle and D.P.W. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1.

[19] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech communication*, vol. 34, no. 3, pp. 267–285, 2001.

[20] C.M. Wessinger, M.H. Buonocore, C.L. Kussmaul, and G.R. Mangun, "Tonotopy in human auditory cortex examined with functional magnetic resonance imaging," *Human Brain Mapping*, vol. 5, no. 1, pp. 18–25, 1997.

[21] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "Self-organizing map in matlab: the som toolbox," in *Proceedings of the Matlab DSP Conference*, 1999, vol. 99, pp. 16–17.

[22] Robert Gütig and Haim Sompolinsky, "The tempotron: a neuron that learns spike timing-based decisions," *Nature Neuroscience*, vol. 9, no. 3, pp. 420–428, Feb. 2006.

[23] Robert Gütig and Haim Sompolinsky, "Time-Warp-Invariant Neuronal Processing," *PLoS Biol*, vol. 7, no. 7, pp. e1000141, 07 2009.

[24] S. Nakamura, et al., "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. ICLRE*, 2000, pp. 965–968.

[25] J. Barker, "Respite CASA toolkit (CTK) v1.3.5," Online, 2004, http://staffwww.dcs.shef.ac.uk/people/J.Barker/ctk.html.

[26] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.