university of groningen

faculty of science and engineering

Keyword Spotting with the Time Difference Encoder

Ton Juny Pina

university of groningen

faculty of science and engineering

University of Groningen

Keyword Spotting with the Time Difference Encoder

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Physics
at University of Groningen under the supervision of
Prof. dr. Elisabetta Chicca (Zernike Institute for Advanced Materials, University of Groningen)
Prof. dr. Beatriz Noheda (Zernike Institute for Advanced Materials, University of Groningen)
Dr. Lyes Khacef (Zernike Institute for Advanced Materials, University of Groningen)
and
Michele Mastella (Zernike Institute for Advanced Materials, University of Groningen)

Ton Juny Pina (s3896781)

December 17, 2021

Contents

Page

Acknowledgments

I want to thank the whole Bio-inspired Circuits and Systems team, for making me feel part of the group since the first moment, discovering me an amazing research field as it is the neuromorphic computing, the unconditional help given daily, the amazing working environment, and just for the very nice people that each of you are. Special thanks to my daily supervisors, Michele and Lyes. Without your guidance and contributions this thesis would not have been possible. I also want to make a special mention to Elisabetta, the mind behind many of the ideas researched in this work. You taught me a lot during this year and it has been a great experience learning and working with you. Finally, I want to thank dr. Beatriz Noheda for also supervising this thesis.

Last but not least, a huge thank you to the family and friends, who has been supporting me during this year and my during my whole life. Without you this thesis would not be possible.

Abstract

In this thesis, the contribution of introducing the Time Difference Encoder to neuromorphic keyword spotting models is researched. Previous studies show that implementing a layer of this neural structure into these models, can improve the performance in the classification task. The reasons for this improvement are discussed by analyzing how the speech is encoded and processed in the biological nervous systems. Information Theory is used in order to research the stimulus encoding in the different parts of the models. In the first set of experiments, a realistic model based on a Python implementation of a biological cochlea is bench-marked, showing successful results by the TDEs in the encoding of temporal patterns from the cochlea representation of the human speech. The second set of experiments tests a model that uses the extracted formants from human speech to classify the spoken words. As in the first experiment, the TDEs show successful results in the encoding of temporal patterns in the formant shape. Finally, a simple binary classifier is designed that performs the keyword spotting from the formants model outputs, showing a solid performance.

Figure 1: Little girl talking to Amazon Alexa. (Shutterstock 2021)

1 Introduction

1.1 Context and motivation

In speech processing, *keyword spotting* deals with the identification of specific *keywords* in spoken language, which can be anything from simple words to full sentences. Performing speech recognition is computationally expensive, due to the model complexity needed to be able to recognize thousands of different words in a speech. Furthermore, keyword spotting models only need to be trained to identify a single expression. This allows for greater model simplicity, which drastically reduces the required number of calculations and lowers the power consumption.

As the number of devices using voice assistants increase, the use of keyword spotting models has broadly extended. By default, speech recognition in the voice assistants is in standby, and only the keyword spotting runs in the background. This prevents the voice assistant from draining the battery of the device, and also prevents false commands. When the keyword is spoken, the assistant 'wakes up' and the speech recognition is activated.

Due to the need for an integrated an power efficient solution, the development of keyword spotting (KS) models in neuromorphic hardware has been explored in recent years[1][2], and its improved efficiency in front of the models in digital hardware has been proven[3]. Nevertheless, these models still face difficulties in distinguishing some words, especially if they show the same vowel structure. Within the aim of improving the accuracy of the neuromorphic KS models, this work explores the idea of introducing a layer of Time Difference Encoders (TDEs), a neural structure that captures the time difference between two pre-synaptic spike inputs.

In order to benchmark the improvement in the classification task, Information Theory is used to explore the stimulus encoding in each part of the model. Also, the reasons why the KS models benefit from introducing this neural structure are discussed. Moreover, introducing a layer of TDEs increases the model complexity, thus increasing its energy consumption. Measurements are performed in order to optimize the number of TDEs, pursuing the best trade between accuracy and energy consumption. Finally, this work hopes to contribute in the understanding of how the information is encoded and processed in the auditory channel.

This thesis follows up on the results obtained by Tobias Mikkuta (University of Bielefeld) in his Master's thesis[4], where the addition of the TDEs to a bio-realistic KS model was proven to increase the true positive rate in the classification task by a 20%.

1.2 The neuromorphic approach

The biological nervous systems operate under extreme small-size and low-power constraints, and its performance cannot be challenged by any man-made processor today. This performance in terms of real-time processing and energy efficiency is due to its computing architecture. Any biological nervous system is constituted by neurons, which are highly parallel and redundant elements that are intrinsically sluggish, noisy and unreliable. Every neuron has in average 10000 synapses that connect it within other neurons, and from them receives impulses through its dendrites, which act as input channels. Then, if the potential inside the neuron reaches a certain threshold, the neuron is activated and an impulse is emitted through its axon. This leads to a way of computing information where memory and processing are shared and influencing each other, radically different from the digital processors based in the von Neumann architecture.

During the last decades, many efforts have been put into mimicking the properties of the very successful biological nervous systems. This has given birth to the field of artificial intelligence (AI), which aims to build models that solve a task by processing the information through an artificial neural network (ANN). These networks are constituted by artificial neurons, which are an abstraction of its biological counterpart. The artificial neuron is represented by a mathematical function that takes a finite number of inputs, and produces an output as a response. Every input is separately weighted, and then its sum is operated through a nonlinear function called the response function, which determines if the neuron is activated. The output of the response function then becomes one of the inputs for a neuron in the next layer.

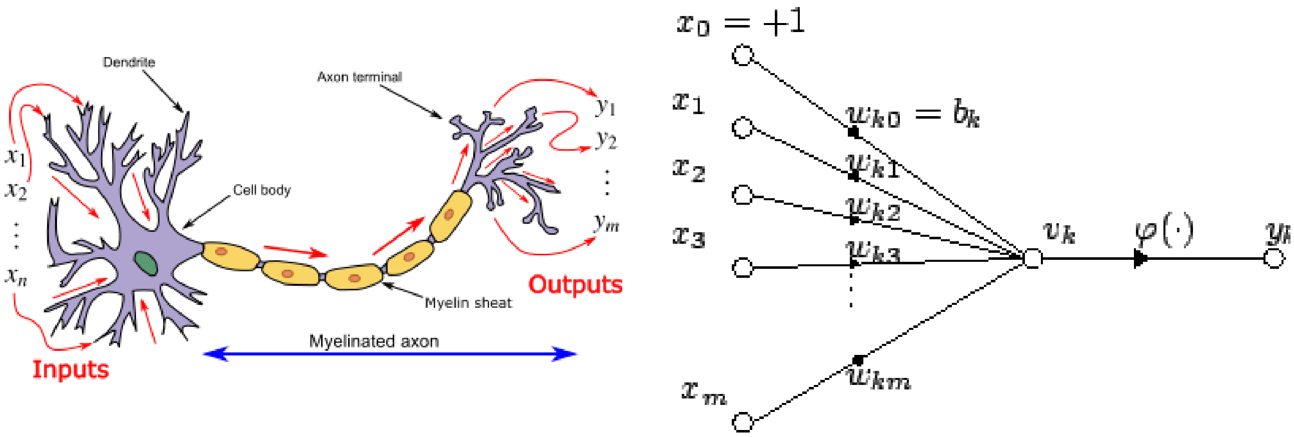


Figure 2: Biological neuron and its mathematical representation, which constitutes an artificial neuron. The n input signals are received in the dendrites. Every dendrite weights the signal by regulating the number of synaptic neurotransmitters. Then the cell body, or soma, adds up all the inputs. If the membrane potential reaches the threshold value, an action potential is sent through the axon to the postsynaptic neurons. In the artificial neuron, this is represented by the response function. (Wikimedia Commons, 2021)

The ANNs can be totally abstract, when they are mathematically computed in a digital processor. In this case, all the synapses can be precisely configured, and there is no limit in the architecture of the network, number of neurons or synapses, except for the computational power required to perform all the calculations. However, this advantage of the simulated networks is also its drawback. As the number of neurons and synapses increases, the number of calculations required to run the model grows exponentially, which in turn increases the power consumption and the computational requirements. For instance, the Human Brain Project[5], which tries to simulate a human-scale cortical model of 20 billion neurons, will require an exascale supercomputer (10^{18} flops) powered with the energy equivalent to a quarter-million households (0.5 GW). The human brain makes it with just 20 W; 50 million times less energy.

In order to move towards the real-time performance and energy efficiency of the biological nervous systems, the field of neuromorphic engineering emulates the dynamics of the neural networks on analogous physical substrate. The neural networks are built from electrical circuits that resemble the properties of the biological neurons. These circuits are built on analog CMOS designed to operate in subthreshold regime, which run on currents of the order of the nanoampere. This greatly increases the efficiency compared to the simulations on digital CPUs, where it's circuits typically operate on currents of the order of the milliampere.

The real-time performance is also improved by the networks in neuromorphic hardware. In standard digital simulations, all the calculations are performed by the CPU, and there is a constant exchange of data between the CPU and the memory. In consequence, the latency in this exchange also constrains the overall performance of the network (which is known as the Von Neumann bottleneck problem). On the other hand, in the neuromorphic networks the information processing is done in parallel by each neuron, as it happens in biology. This leads to producing the outcomes of the network in real-time with very small latency times.

1.3 Research questions

To summarize, this thesis focuses on the following problems:

Q1. Proof that introducing a layer of TDEs improves the classification task in keyword spotting models.

Q2. Explore how the TDEs contribute to keyword spotting. Which are the sound features captured by the TDEs that allow for the stimulus classification.

Q3. Model optimization. Find which TDEs are carrying the relevant information for detecting a certain keyword.

Q4. Explore the information encoding and processing in the auditory channel.

2 Methods

2.1 The Time Difference Encoder

The Time Difference Encoder (TDE) is a neural structure introduced by Milde et. al. in 2018[6]. This neuron receives two presynaptic inputs, and encodes the time difference between them into a burst of spikes. Both the number of spikes and the inter-spike interval are proportional to the time difference between the inputs.

This neural structure is constituted by a leaky-integrate-and-fire neuron and two synapses, the facilitatory synapse (Figure 3a, red) and the trigger synapse (Figure 3a, blue). When a spike is received in the facilitatory synapse, an excitatory post-synaptic current (EPSC) rises the TDE membrane potential. As shown in figure 3b, the synaptic weight of the facilitatory synapse is set so the membrane voltage does not reach its threshold value with the EPSC from the facilitatory synapse. The efficacy of the trigger synapse is determined by the EPSC in the facilitatory synapse. If a spike is received while the EPSC from the facilitatory synapse is non-negligible, the current from the trigger synapse rises the membrane potential over the threshold and the TDE spikes. The current integrated by the neuron is then proportional to the time difference between the presynaptic spikes, thus its number of spikes (Figure 3e).

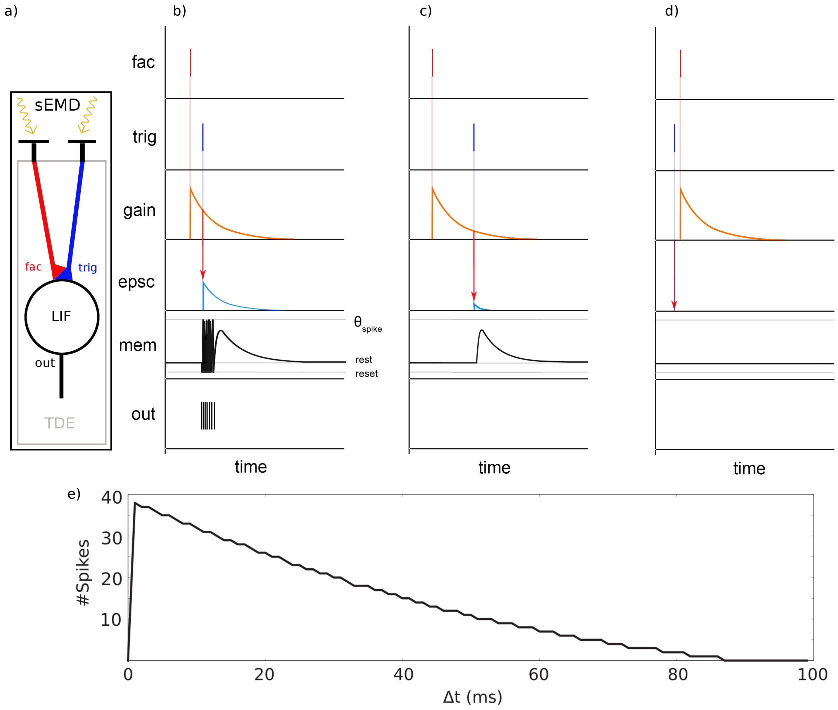


Figure 3: Representation of the TDE response to the input spikes. In 3a the TDE neuron is represented with its facilitatory synapse (red) and its trigger synapse (blue). Following figures show the TDE response to spikes in the correct order and short time difference (3b), long time difference (3c) and anti-preferred order (3d). Finally, figure 3e shows the number of spikes in the TDE response as a function of the time difference between facilitatory and trigger inputs. (D'Angelo et al., 2020)

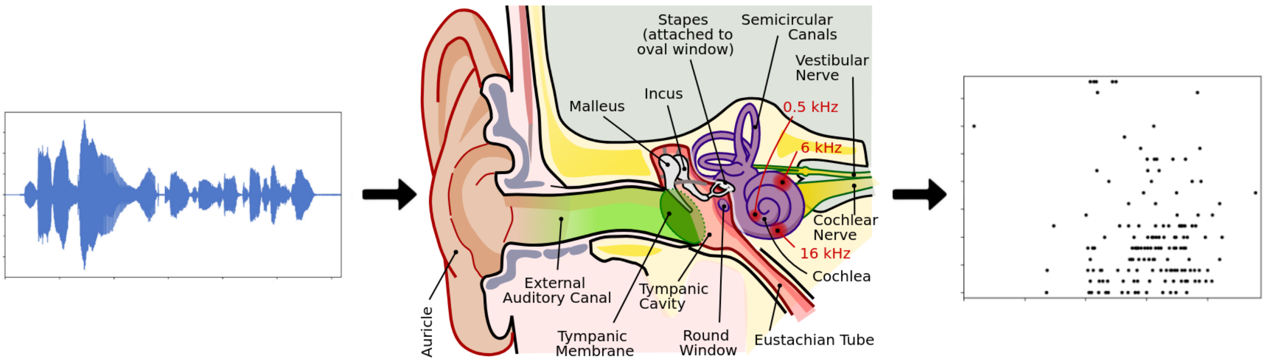


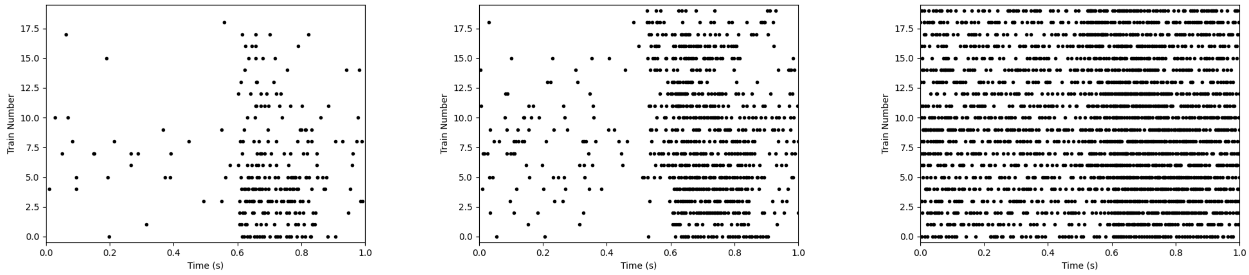
Figure 4: Representation of the speech processing in the auditory channel. The sound waves received in the inner ear are encoded by the spiking patterns of the fibers in the cochlea. (Chittka and Brokmann, 2005)

The TDE model used in this research is a Python implementation in Nengo[8], provided by Terry Stewart.

2.2 Speech processing

In order to study the contribution of using TDEs in keyword spotting, some understanding about how the speech is encoded and processed in the auditory channel is required. This section gives some insights about the characteristic properties of a sound-wave that relate it to the uttered word, and how this are encoded by the TDEs.

The cochlea is a hollow, spiral-shaped bone found in the inner ear. When the sound is processed the auditory channel, it produces vibrations in the cochlea. This vibrations are encoded into electrical impulses that correspond to the sound amplitude for each individual frequency. Mathematically, it acts as a discrete Fourier transform over the sound wave



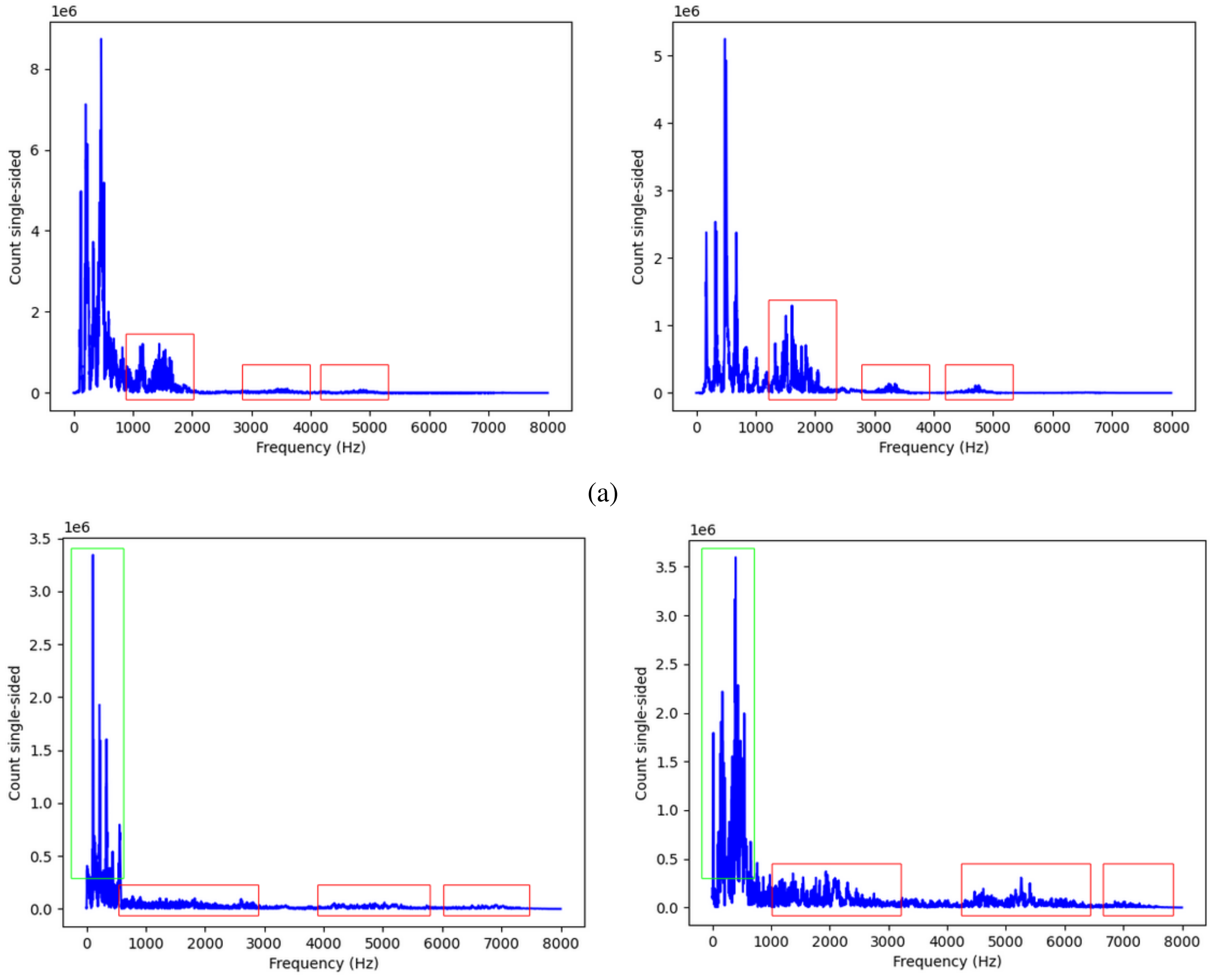
(a) (b) (c)

Figure 5: Comparison of the resulting spike-trains from processing the same utterance of the word *two* with 20 equally spaced channels between 0 Hz and 8000 Hz, and 1 LSR (a)/MSR (b)/HSR (c) fiber per channel[10]

$$f(t) = \sum_{n=0}^{N} A_\nu \exp\left(-i\frac{2\pi\nu_n}{N}t\right), \tag{1}$$

where A_ν represents the amplitude for a certain frequency ν, and the sum goes over all the discrete frequencies. This work uses the Python implementation of a bio-inspired cochlea model developed by Zilany et al. in 2009[10][11].

The neurons that encode the amplitude in each frequency are called fibers. The encoding is done by the spike-rate. Fibers in biological cochleas show a wide range of sensitivity, which improves the hearing capabilities in the real world. It allows our brains to perform well in noisy environments, to modulate the sensitivity regarding the sound volume around the subject and to differentiate between sounds in situations with many sound sources. This is represented in our model by three types of fibers: the High Spontaneous Rate (HSR), the Medium Spontaneous Rate (MSR) and the Low Spontaneous Rate (LSR) fibers (figure 5).



(a)

(b)

Figure 6: Accumulated sound amplitude per frequency during two utterances of the word *one* (a) and *two* (b).

The human speech is produced by modulating the flow of air expelled from the lungs through the voice tract. This produces vibrations of different frequencies that constitute each phoneme. The vowel sounds are the easier to identify, since are characterized by high peaks of amplitude in the low frequencies range (200 Hz -2 kHz), which are produced by strong vibrations in the vocal chords. The consonant sounds are usually characterized by fainter peaks of amplitude through the whole frequency range, result of modulations made by the throat, mouth, tongue and lips.

In the figure 6, the sound amplitude per frequency for two utterances of the words *one* and *two* is shown. Regardless of the speaker, the structure of peaks is common in different instances of the same word, even the center position of the peaks may vary depending on the voice tone of the speaker. Every word, or more precisely every one of the phonemes that form a certain word, are defined by a superposition of sound-waves of certain frequencies. This phonemes are encoded by the cochlea as a subset of channels spiking simultaneously. The temporal evolution of the spiking channels generates characteristic patterns for each word (figure 7).

The TDEs aim to extract this patterns in the temporal evolution of the spike-trains from the cochlea, and encode them into the activation of certain TDEs. Since these patterns are found to be common in instances of different speakers pronouncing the same word, for a certain word the most active subset of TDEs and its sequence of activation identifies the utterance.

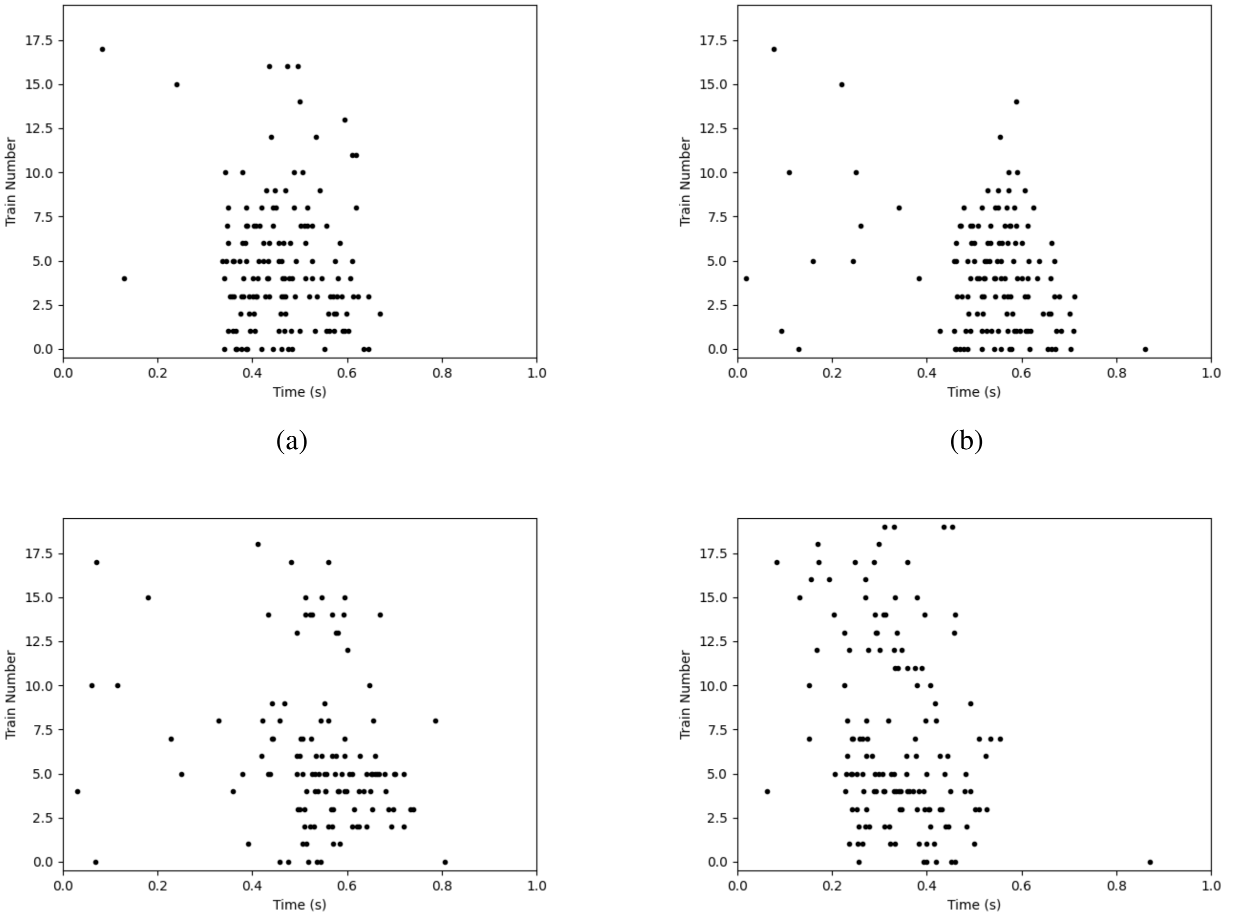
The higher amplitude peaks in the frequency domain are the formants. Their spatio-temporal evolution can also be used to characterize the corresponding utterance, as a simplified version of the patterns observed in the cochlea spikes. The position of the first two formants is enough to identify a vowel sound, and the four first formants give enough information for identifying most of consonant sounds[add ref. code formants extraction]. Comparing the cochlea outputs with the corresponding formants for *one* and *two* (figure 8), a high density of spikes is observed coinciding with the position of the formants, creating patterns that resemble the shape of the formant.

2.3 Information theory: Entropy and Mutual Information

In order to test the performance of the models, and study how the stimulus are encoded and processed, this work uses Information Theory.

Information theory is a useful analysis tool for neuroscience data. It is model independent, so it is not necessary to hypothesize a specific structure to the interactions between variables. The model-free character allows a much wider range of interactions to be quantified than could be achieved with a model-dependent approach, that is limited by the assumed model. Information Theory can be applied to any mixture of data types, which is really helpful when comparing stimulus presented to a network model with the spiking outputs. It is also capable of detecting non-linear interactions. Given the prevalence of non-linear interactions in neural networks, this ability is especially important. Finally, the results from the information theoretic measurements are in the general unit of bits, which facilitates comparisons between results.

In Information Theory, the information is defined as the reduction in the uncertainty about the state of a variable when the state of a second is known. To illustrate this concept, imagine that the reader unexpectedly encounters with the author of this thesis. The author flips a coin and hides the result.



(a)

(b)

(c)

(d)

Figure 7: Resulting spike-trains from two utterances of the word *one* (a,b) and *two* (c,d). The cochlea is configured with 20 equally spaced channels between 0 Hz and 8000 Hz and 1 LSR fiber per channel.

The reader asks the question: ”Did it come up tails?”, which is trustfully answered: ”Yes”. The answer totally reduced the reader’s uncertainty about the state of the coin. Since the coin had two equally probable states, the author’s message contained 1 bit of information.

Moving from coins to the context of spiking data, one might try to ask questions about how much information a spike train (analogous to the answer about the state of the coin) provides about a stimulus (analogous to the result of the coin flip). In this case, asking Yes/No questions is not enough to provide information measurements. For this purpose, information theory provides useful quantities that can be calculated.

The first quantity that needs to be introduced is the entropy of a variable. Entropy is the fundamental IT quantity, and it measures how much uncertainty is contained in a variable. The entropy H(X) of a random variable X (with individual states x) is defined as

$$H(X) = \sum_{x \in X} p(x) log \left(\frac{1}{p(x)} \right), \tag{2}$$

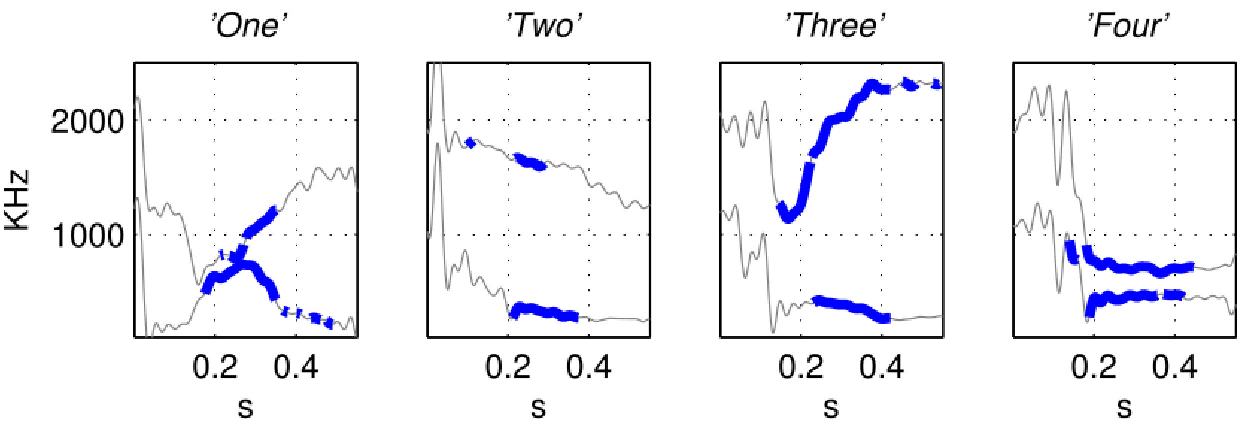
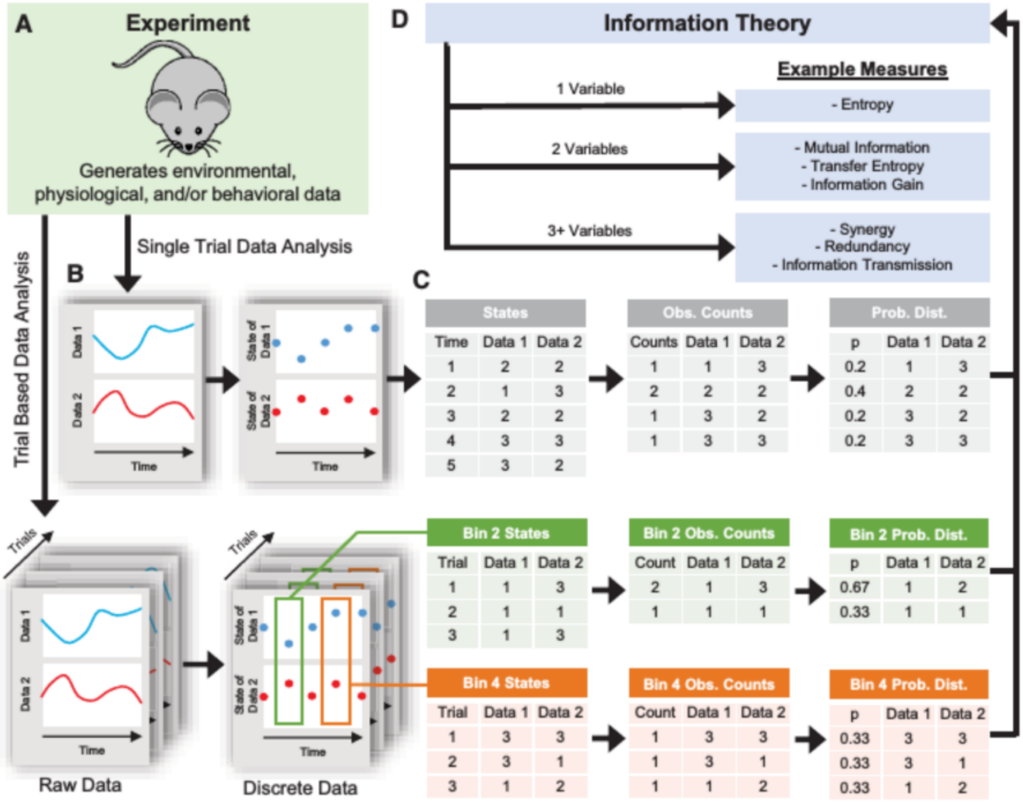


Figure 8: First and second formants extracted from utterances of the words *one*, *two*, *three* and *four*. (Coath et al., 2014)



States

Time	Data 1	Data 2
1	2	2
2	1	3
3	2	2
4	3	3
5	3	2

Obs. Counts

Counts	Data 1	Data 2
1	1	3
2	2	2
1	3	2
1	3	3

Prob. Dist.

p	Data 1	Data 2
0.2	1	3
0.4	2	2
0.2	3	2
0.2	3	3

Bin 2 States

Trial	Data 1	Data 2
1	1	3
2	1	1
3	1	3

Bin 2 Obs. Counts

Count	Data 1	Data 2
2	1	3
1	1	1

Bin 2 Prob. Dist.

p	Data 1	Data 2
0.67	1	2
0.33	1	1

Bin 4 States

Trial	Data 1	Data 2
1	3	3
2	3	1
3	1	2

Bin 4 Obs. Counts

Count	Data 1	Data 2
1	3	3
1	3	1
1	1	2

Bin 4 Prob. Dist.

p	Data 1	Data 2
0.33	3	3
0.33	3	1
0.33	1	2

Figure 9: Graphic representation of the Information Theory processing of experimental results. (Timme and Lapish, 2018)

where the sum is over all the possible states of the variable X, and it is measured in bits. Then in order to calculate the entropy of a variable, the probability distribution over its possible states needs to be estimated by experimental measurements.

The entropy definition can be generalized to two (or many) variables by substituting the probability distribution $p(x)$ for the joint probability distribution $p(x,y)$ in 2,

$$H(X,Y) = \sum_{x\in X, y\in Y} p(x,y) log\left(\frac{1}{p(x,y)}\right) \tag{3}$$

which is the joint entropy of X and Y.

In the case of independent variables, the joint probability distribution is just the product of both variables distributions, then the joint entropy

$$p(x,y) = p(x)p(y) \Rightarrow H(x,y) = H(x) + H(y) \tag{4}$$

is just the sum of the individual entropies for each variable.

The entropy reduction in a variable due to the knowledge about the state of a second is quantified by the conditional entropy

$$H(X|Y) = \sum_{x\in X, y\in Y} p(x,y) log\left(\frac{1}{p(x|y)}\right). \tag{5}$$

$p(x|y)$ refers to the probability distribution of X once known the state of Y, and as in the previous cases the sum is over all the possible combinations of states for X and Y.

The information has been defined as the reduction of uncertainty in the state of one variable when another is known. So from the definitions of entropy and conditional entropy, it is straight forward to define the mutual information between to variables as

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in X, y\in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{6}$$

By performing mutual information measurements on the spike trains before and after the TDE layer, the models performances and the stimulus encoding by the TDEs is studied. The calculations are performed with the toolbox provided by [13], a Python package for Information Theoretic analysis of neural data.

2.4 Bio-inspired model

The first set of experiments tests the network from T. Mikkuta's Master's thesis[4]. The model takes a biologically realistic approach, by using the Python implementation of a biological cochlea [10]. The cochlea is fed with audio samples from the *Google speech commands* dataset, which includes 65000 one-second long utterances of 30 short words recorded by thousands of different speakers.

The outputs of the cochlea are processed by a layer of TDE neurons, which are implemented in Nengo[8]. Each TDE takes the spikes from one cochlea channel in its facilitatory synapse, and the spikes from another channel in its trigger synapse. Thus, the TDEs are encoding the time differences between spikes in different channels in the cochlea. There is a TDE for all the possible combinations of cochlea channels as facilitatory and trigger inputs. In consequence, if N is the number of channels in the cochlea, the number of TDEs in this layer scales as N^2.

This scaling factor in the number of TDEs is not very efficient in terms of energy consumption, but the hypothesis is that only a reduced population of TDEs is actually needed in order to capture the characteristic spatio-temporal patterns that identify a certain keyword. If this TDEs can be identified, the keyword spotting task can be performed by using only a few neurons and synapses, achieving great energetic efficiency.

In the research from T. Mikkuta[4], the spiking outputs from the cochlea and the TDEs were fed to a least-squares word solver, which performed the classification task for the keyword spotting. Its results were then compared, observing an improvement in the True Positive Rate (TPR) and the accuracy when using the spike-trains from the TDE layer. In this work, Information Theoretic measurements are used to study and compare the stimulus encoding in each layer of the model. The aim of the experiments is to verify the improvement provided by the TDEs, and to explore if reduced populations of TDEs can be used to capture the key features in a spoken word and perform the keyword spotting task efficiently.

The hyper-parameter settings for the TDE layer are taken from T. Mikkuta research[4]. This values were found to provide the best accuracy performing the KS task with this model. Even though further studies should be performed in order to find more accurate values.

2.4.1 Experiment 1: Mutual Information in the cochlea and TDEs

The first experiment explores the information encoding in the cochlea and TDE layers. When a stimulus is encoded into a sequence of spikes by a group of neurons, the information can be encoded in different ways. If it is the timing of the spikes and its frequency, the features that encode which stimulus is presented to the neurons, the encoding is temporal. On the other hand, if the feature that encodes the information about the stimulus presented to the neurons is which neuron spikes, the encoding is spatial.

Furthermore, this experiment aims to prove that the information is encoded more spatially in the TDE layer than in the cochlea. Most of the information about the stimulus is encoded by temporal patterns in the cochlea, by the spikes timing and rate in each channel which corresponds to the sound amplitude in a specific frequency range. In the next layer, the TDEs encode the stimulus by capturing the time differences between spikes in different channels. When the stimulus is encoded in the TDEs, part of this information represented by temporal features in the cochlea is translated into the activation of certain TDEs (i.e. it is spatially distributed).

In order to estimate the probability distributions for each stimulus in the calculation of the mutual information, the algorithm needs to compare the obtained results over all its possible outcomes. This translates to $2^{\alpha\beta}$ possible matrices for each spike-train, where α is the number of channels and β is the number of time-steps in the simulation. In order to reduce the number of possible outcomes, so it is feasible the computation with a desk CPU, the spike-train matrices are simplified before computing the mutual information.

Following the purpose of this experiment, which is to compare the spatially encoded information in the two layers of the model, the simplification eliminates the temporally encoded information in the spike-trains. This is done by considering only the total number of spikes in each channel in the mutual information calculation. With this simplification, the number of possible outcomes is α^{β}, where α is the highest spike-count and β is the number of channels in the layer. This is still not manageable by a CPU, so further simplification is needed.

The final simplification is done by following a winner-take-all strategy, where the most spiking channel is selected from the total spike counts. Then the mutual information is computed between the most spiking channel in the cochlea/TDE layer, and the stimulus presented. This simplification reduces the number of possible outcomes to the number of channels of each layer.

The dataset employed in the experiment consists of 500 instances of the words *zero*, *one*, *two* and *three*. The whole dataset is used in the calculation of the probability distribution of the response to each stimulus, which guarantees low bias in the MI calculation[14].

2.4.2 Mutual Information in reduced populations of TDEs

The second experiment aims to explore which TDE neurons are encoding more information about the stimulus. For this purpose, the TDE neurons have been divided into different populations regarding the frequency of the cochlea channels that are receiving inputs from. For instance, the 'low' population consists of the TDEs that receive both inputs (facilitatory and trigger) from low-frequency channels in the cochlea; the 'low-high' population receives one input from a low frequency channel and the other from a high-frequency channel, etc.

The used dataset is composed by 200 instances of the words *zero* and *one*. The mutual information is calculated from the spike-trains of a reduced number of TDEs in each population, which are randomly selected. In order to make the calculations feasible, the possible outcomes are reduced through a binning method. The binning is performed by adding the number of spikes inside a defined time-bin, and setting the value for that time-bin into 1 if it's above a certain threshold, and 0 if it's below. In this experiment, the aim is to evaluate which TDE neurons are encoding more information about the stimulus presented, both if its temporally or spatially encoded. To this end, this method intends to maintain the temporally encoded information, unlike the simplification method used in the first experiment.

2.5 Formants model

In the section 2.2, the formants are introduced as the local maximums in amplitude in the frequency domain. Vowel sounds have between 4 and 6 distinguishable formants in the low frequency range (0 - 2.5 kHz). Moreover, the positions of the two formants with the lowest amplitude, which are the two first formants, give enough information to distinguish between vowel sounds[15].

In previous research, a neuromorphic keyword spotting model has been developed[1] which successfully uses the extracted formants from spoken words to perform the classification task. Following its results, the performance of a TDE-based keyword spotting model that uses the extracted formants for the classification task is assessed.

The formants are extracted from audio samples of the *Google speech commands* dataset using Sinewave Speech analysis[16]. Based on the LPC speech analysis, this algorithm tracks the frequencies and amplitudes of the four first formants by fitting sinusoids waves to the speech sample. The positions of the four first formants are stored as 32 x *n* binary patterns, with each of the 32 rows representing a frequency channel and each of the *n* columns representing a time bin of 1 ms (Figure X). By these procedure, a dataset consisting of 500 formants of the spoken numbers from 0 to 9 is created. These are used as inputs to the network simulation.

The network is composed by a single layer of TDEs implemented in Nengo[8]. In the bio-inspired cochlea model, the TDE layer contains a TDE for all the possible combinations of channels as facilitatory and trigger inputs. In this case, the number of TDEs have been reduced by defining a maximum distance D_{max} between synapses in the TDEs. The distance between two channels is defined as the difference between its channel numbers, i.e. channels 2 and 5 are at a distance of 3 channels. Then, there is a TDE for all the possible combinations of channels at a distance equal or smaller than D_{max}.

The benchmarking of this model aims to validate the hypothesis that capturing the temporal evolution of each of the first four formants with the TDEs, captures enough information about the stimulus to perform the keyword spotting task. In line with this purpose, D_{max} is defined in order to avoid that the TDEs receive inputs from two different formants in its facilitatory and trigger synapses. After visually assessing some of the extracted formants, the maximum distance D_{max} is set as 3 channels. This value is found to be in most cases small enough to avoid taking inputs from different formants in the TDEs, while being at the same time large enough to capture the temporal evolution of the formant shape. Setting $D_{max} = 3$ leads to a total of 182 TDEs in the network.

2.5.1 Experiment 1: MI reduced populations

The aim of this experiment is to find the best configuration for the TDE layer hyper-parameters. To this aim, a brute-force approximation is carried out. In order to reduce the number of parameters, the weights of all the TDEs are set uniformly, and with equal value for both synapses $\omega_{fac} = \omega_{trig} = \omega$. Also the decaying constants for the EPSC are set uniformly in all the TDEs, but with different value for each synapse ($\tau_{fac} \neq \tau_{trig}$). Then, the experiment is repeated for the 60 possible combinations with $\omega = (20000, 40000, 50000, 60000, 80000)$, $\tau_{fac} = (0.002$ s, 0.003 s, 0.005 s, 0.008 s, 0.012 s, 0.015 s) and $\tau_{trig} = (0.001$ s, 0.002 s).

The experiment consists in two phases. In the first phase, which is defined as the training phase, a keyword is selected from the dataset. Then, 100 formants of the keyword are used as inputs for the TDE layer, and the outputs are stored. After storing the formants and the TDE layer outputs, the number of spikes in each formant channel and in each TDE through the whole training phase are computed. From the total spike counts, the channels in each layer regarding its overall spikecount in the training phase are ranked.

In the second phase of the experiment, defined as the testing phase, 200 formants of each word in the dataset are used as inputs for the TDE layer, and its outputs are stored. For each trial, the spike counts in a reduced population of TDEs are computed, as well as in a reduced population of the formant channels. The reduced populations are defined as different percentages (from 5% to 40%) of the top ranked formant channels/TDEs regarding its overall spikecount in the training phase. Then, the mutual information is calculated between the total spikecount in each population of formant channels/TDEs and the stimulus presented (keyword or not-keyword).

The MI values are represented as a function of the number of formant channels/TDEs, which allows to evaluate the populations needed to capture enough information about the stimulus. The results through the 60 trials are also visually assessed in order to find the best performing configuration for the TDE layer.

2.5.2 Experiment 2: Spikecount-based classifier

The second experiment tests the performance of a classifier performing the keyword spotting task with the formants. The aim of the experiment is to compare the results when the classifier is using the formants for the stimulus classification, and when is using instead the TDE layer outputs. The hyper-parameter configuration is set as the best performing configuration found in the Experiment 1.

As in the previous experiment, this experiment also consists in two phases. In the training phase, 400 formants of the selected keyword are fed to the TDE layer, and its outputs are stored. Then, the total spike counts are calculated for each formant channel and for each TDE. From the total spike counts, the formant channels and the TDEs are ranked regarding its overall spikecount. For a reduced population, consisting in the X% of the top ranked formant channels and top ranked TDEs, the mean spike counts per trial and its standard deviation are also computed.

In the classification phase, 100 formants of each word in the dataset are used as the testing dataset. The formants are fed to the TDE layer, and its outputs are stored. In each trial, the classification of the stimulus is made from the spike counts in the most spiking X% of the formant channels and TDEs. A threshold θ is calculated from the mean spike counts in the training phase of the top ranked population, and its standard deviation. If the spike count in top ranked channels for a certain trial is higher than the threshold, the stimulus is classified as keyword. On the other hand if the spike count is lower, is classified as not-keyword.

In order to assess the classifier performance, the Receiver Operating Characteristic (ROC) curves[17] are computed by ranging the threshold from $\theta = C - 3\sigma_C$ to $\theta = C + 3\sigma_C$, where C are the mean counts per trial in the top ranked population, and σ_C its standard deviation. For each threshold value the confusion matrix is calculated, and also the True Positive ratio (TPR), the False Positive ratio (FPR) and the accuracy.

3 Results and discussion: Bio-inspired model

In the following section, the results from the benchmarking of the bio-inspired model described in the Methods section are presented and discussed. Mutual Information is used to compare the stimulus encoding in each part of the network, with the aim of proving that the addition of the TDEs eases the classification task. Also, the Mutual Information from small populations of TDEs is measured, in order to explore the minimum populations of TDEs required to spot a keyword.

3.1 Mutual Information in the cochlea and TDEs

The aim of this experiment is to serve as a first proof of concept that the stimulus encoding is more spatially distributed in the TDE layer, thus easing the classification task. Figure 10 compares the mutual information values between the spike-trains and the presented stimulus after the spike-train simplification, for both the cochlea and TDE layers. In this sense, figure 10 shows around a 100% increase in the mutual information values for the TDE layer after the temporal information has been removed by the reduction method. In consequence, the TDE layer is effectively encoding the temporal patterns in the cochlea with the activation of certain TDEs.

In the brute-force approximation to the optimal hyper-parameter settings from T. Mikkuta's research[4], the best accuracy was found in the highest values tested. Repeating the comparison of the mutual information in the cochlea versus the TDE layers with MSR fibers, shows higher mutual information values than with LSR fibers (Appendix, figure 16). Even though the mean inter-spike interval in the MSR case is way shorter, the same hyper-parameter settings are used. This results point in the direction that longer values for $\tau_{facilitatory}$ and $\tau_{trigger}$ would improve the model performance with LSR fibers.

3.2 Mutual Information in reduced populations of TDEs

In this experiment, the information encoding in the TDE layer is explored. To this end, Figure 11 compares the mutual information values for different number of TDEs selected in each population. In figure 12a, the mutual information values for different sizes of time-bins are represented. Finally, figure 12b shows the results of introducing a scale factor. This consists in increasing the weights of the trigger synapse as $\omega = f_{\omega} \cdot \#_{ch}$, where f_{ω} is the scale factor. By introducing the scale factor, the weights linearly increase with the frequency of the fiber.

The exploration of which populations of TDE encode more information about the stimulus, follows previous studies about speech processing. The vowel structure of a word is represented by its peaks of amplitude, its formants, in the region of 0-2500 Hz. In this case the 4 words presented have different vowel structure, so the information encoded by the TDEs connected to the low frequency channels may be enough to classify the stimulus. This is seen by the MI values measured in the low population in figures 11 and 12. Also most of the sound amplitude in the human speech accumulates in this range of frequencies, so the population of TDEs connected to the low frequency channels show more activity.

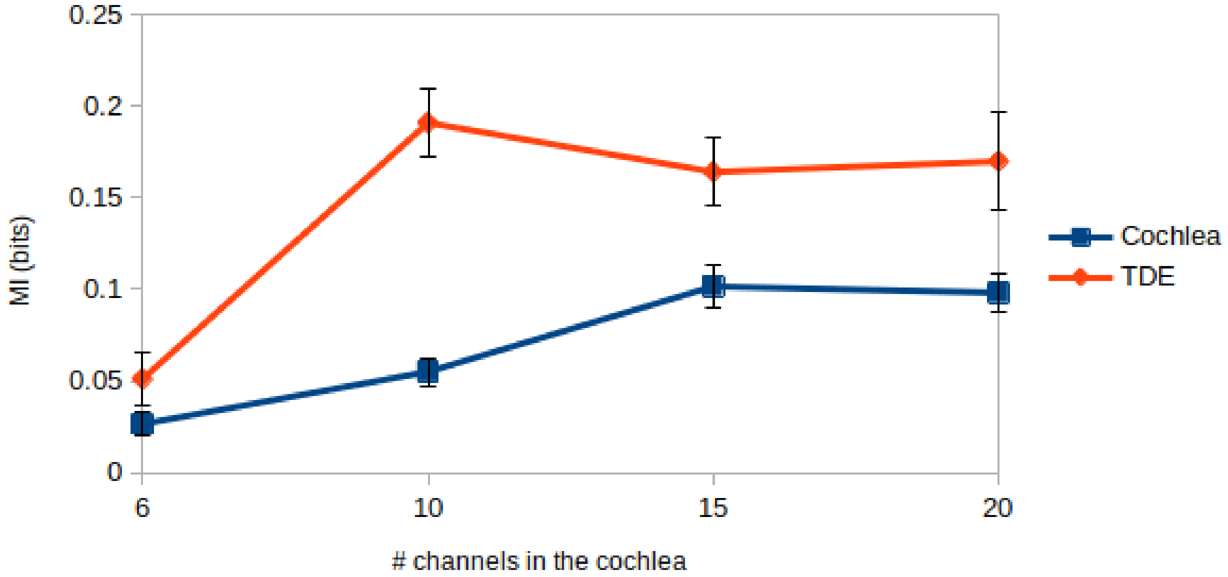


Figure 10: Comparison of the Mutual Information values measured between the stimulus presented to the network and the cochlea spike-trains (blue) or the TDE layer spike-trains (orange), as a function of the number of channels in the cochlea.

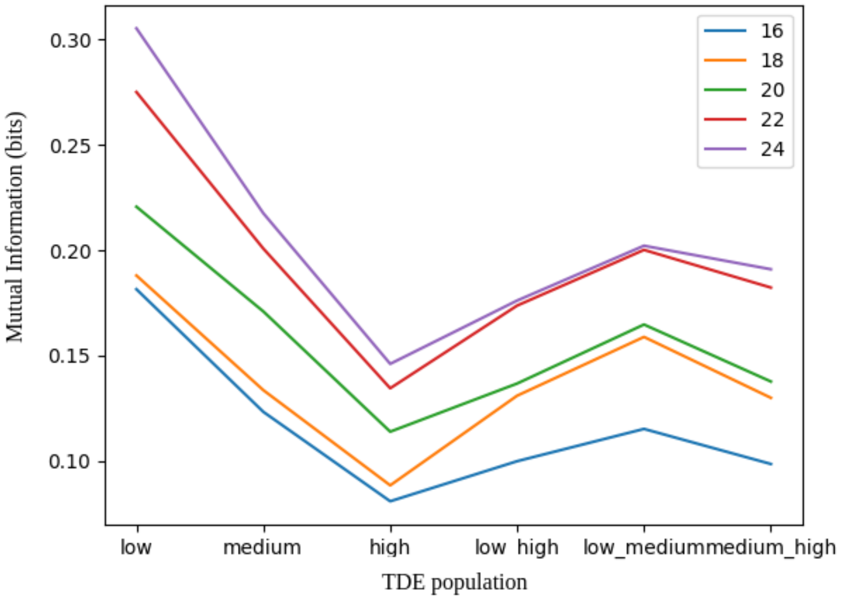
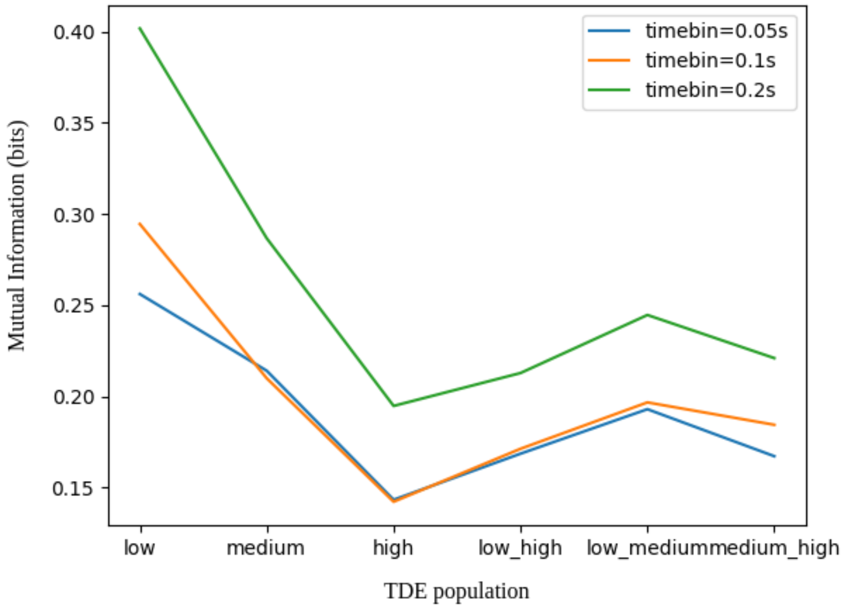
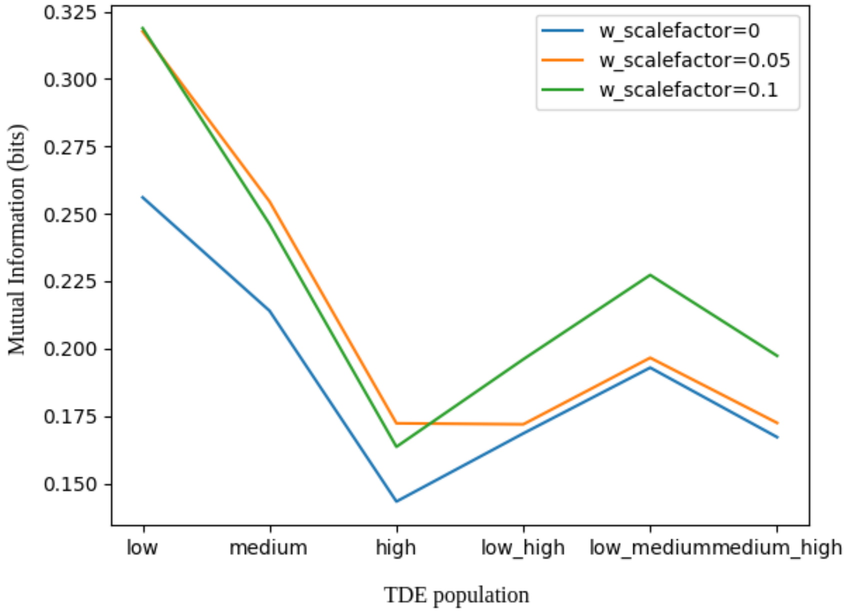


Figure 11: Mutual information that subsets of 16 to 24 TDEs provide about the stimulus, chosen randomly over each of the populations.

In order to improve the accuracy of the keyword spotting model, the goal must be to avoid false positives for words with similar vowel structure. For this purpose, the information encoded by the TDEs receiving inputs from higher frequency populations is important. This is shown by figure 11. In this model, the weights for all the TDEs are set in the same value. As most of the activity in the



(a)



(b)

Figure 12: Mutual information measurements with subsets of 24 TDEs from each population. Figure a) shows the results for different time-bin sizes. In figure b), the results from applying different scaling factors in the weights of the trigger synapses is shown.

cochlea locates in the low frequency channels, the weight's value is set as the TDEs connected to this channels do not saturate. The drawback is that some spikes in high frequency channels, which carry important information about the stimulus, do not trigger any TDE to spike. Introducing the scale factor in the weights, allows the TDEs to capture easily features of the stimulus represented by a only a few spikes in high frequency channels. Thus, figure 12b shows higher MI when the scale factor in the weights is introduced.

It is also important to note that the results shown in the Figure 12a point in the same direction than the previous experiment. The mutual information values for all the TDE populations are higher as the size of the time-bin increases. Thus, it indicates that it is more important the information carried by the number of spikes of each TDE, rather than the precise timing of the spikes, as the precise timing is lost as the size of the time bin increases.

4 Formants model

In the following section, the results from the experiments performed with the formants model described in the Methods section are presented and discussed. The aim of this set of experiments is to test the performance of the TDEs in a simpler and less noisy representation of the stimulus, in terms of assessing if the TDEs are extracting the information about the stimulus that is represented in the temporal evolution of the formant, and encoding it spatially by the spiking of a certain TDE.

4.1 Mutual Information in reduced populations of formant channels and TDEs

In this experiment, a brute-force approximation to the optimal values for the TDE layer hyperparameters is performed. After visually assessing the results of the 60 tested combinations of ω, τ_{fac} and τ_{trig}, the best performance is found for $\omega = 50000$, $\tau_{fac} = 0.008$ s and $\tau_{trig} = 0.002$ s. Figure 13 shows the Mutual Information values both for the formant channels and the TDEs, as a function of the number of formant channels/TDEs included in the spike counts.

Regarding that the MI is calculated only by the spike counts in certain channels and TDEs, all the information about the stimulus that is represented by the timing of the spikes is not taken into account. In this sense, comparing the MI values for the TDEs and for the formant channels show that while in the TDE layer most of the information about the stimulus is represented by which TDE is spiking and its spike count, in the formant the timing of the spikes and the temporal evolution of the formant plays a bigger role. Thus, the TDEs are extracting temporal features from the formant that characterize the stimulus, and encoding them in the activation of certain TDEs.

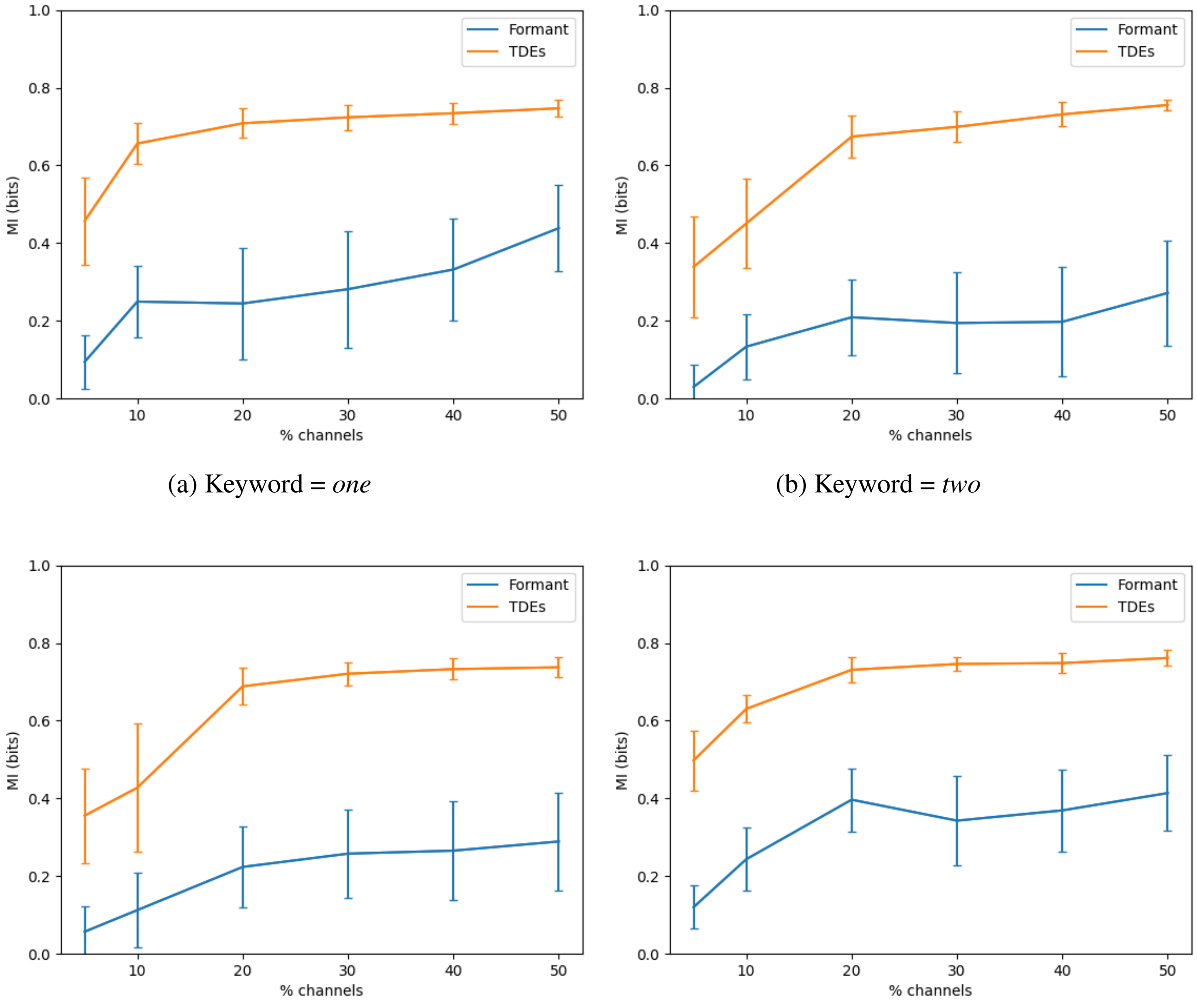
The assessment of how the MI increases as a higher number of formant channels/TDEs is taken into account, shows that using between a 10% and 20% of the TDEs is enough to capture most of the information about the stimulus. This result is very promising in terms of developing energetic efficient keyword spotting models. Regarding that the total number of TDEs in this model is 182, this result means that only with 20-30 TDEs is enough to extract the characteristic temporal features of the keyword that allow its classification with this set up.

4.2 Spike-count based classifier

In this experiment, the performance of a classifier based on the spike counts in a reduced number of channels/TDEs is tested. The channels/TDEs are selected as the most spiking channels/TDEs for the specific keyword in the training phase. The performance is visually assessed by the ROC curves, and also by the True Positive rate (TPR), the False Positive rate (FPR) and the accuracy for the best performing threshold in each ROC curve.

The comparison of the ROC curves for the stimulus classification with the formants and with the outputs of the TDE layer (Figure 14), shows that the classifier performs better with the TDE outputs than with the formants. As in the previous experiment, all the information encoded by timing of the spikes is lost by only taking into account the spike counts. In this sense, the better performance of the TDE layer follows the previous results, as the TDEs are able to spatially encode information represented by the temporal evolution of the formant, thus easing the classification task.

Comparing the ROC curves in Figure 15, is important to note that when only the spike counts of a small number of channels/TDEs is used in the classification (1% - 6%), the classifier is able to hold better its performance when using the TDE layer outputs than when using the formant. This result also points in the direction that the TDEs can be useful for developing a very energetic-efficient keyword



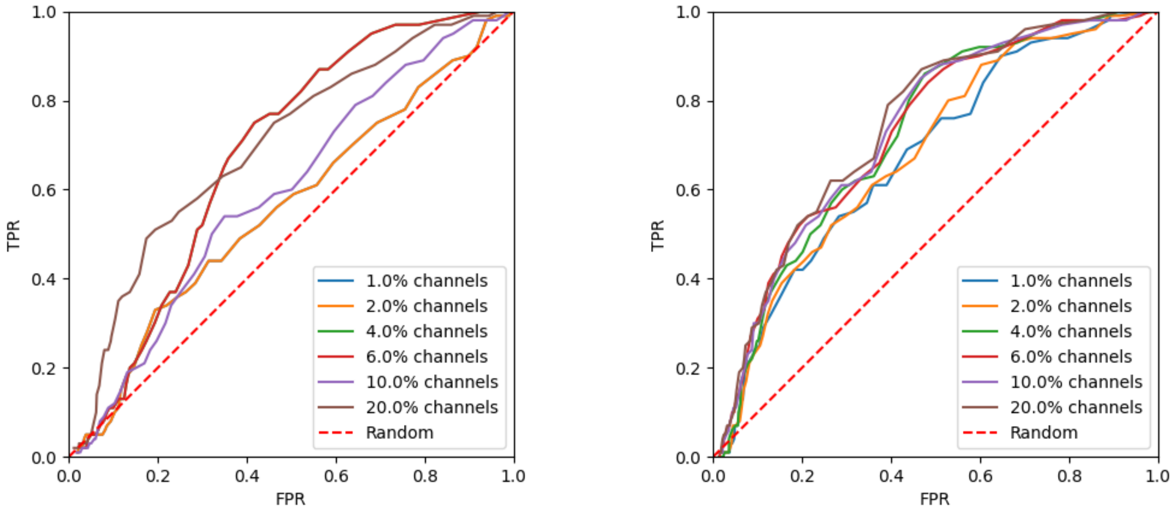
(a) Keyword = *one*

(b) Keyword = *two*

(c) Keyword = *three*

(d) Keyword = *four*

Figure 13: Mutual Information between the spike counts in the most spiking formant channels (blue) and TDEs (orange) and the stimulus presented, as a function of the percentage of the formant channels/TDEs included, for the best performing configuration of the TDE layer hyper-parameters. The channels/TDEs are ranked regarding its total spike count in the training phase, and the populations are chosen regarding the X% of channels/TDEs that showed the higher spike count. The different figures show the results for setting each of the words in the dataset for this experiment as the keyword for the training phase.



(a)

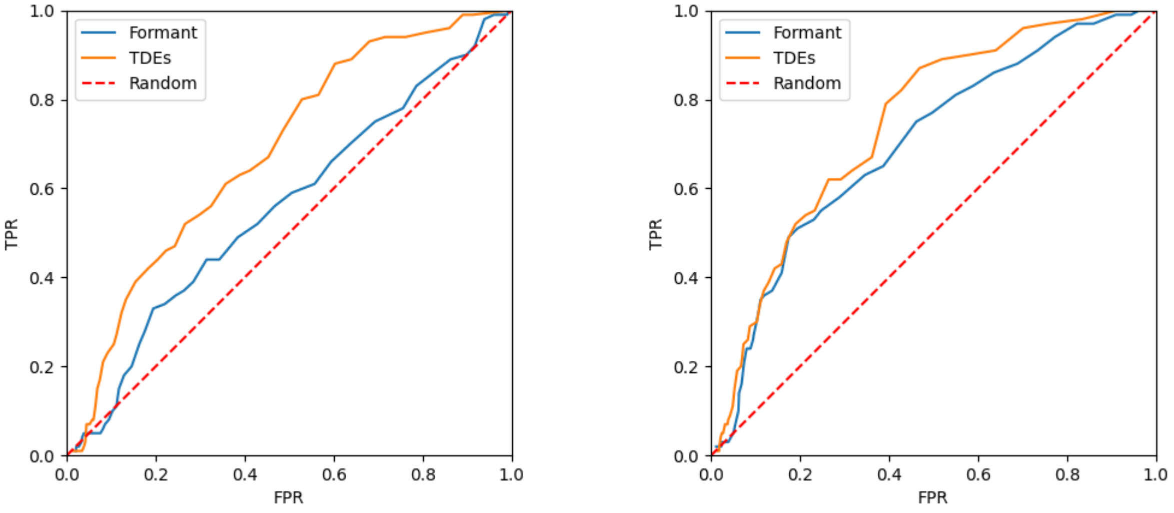
(b)

Figure 14: ROC curves for the classification of the stimulus by the spike counts in the formant channels (a) and by the spike counts in the TDE layer (b) for the keyword *one*. Each curve corresponds to the percentage of the formant channels/TDEs used in the classification of the stimulus, and the red dashed line corresponds to a random classifier.

spotting model with a small number of neurons.

The best performance for the classifier, is obtained from the spike counts of the TDE layer when using the 20% of the TDEs. For the best performing threshold, which is obtained from the values represented in Figure 15, gives an accuracy of 79.8%, with a TPR of 0.71 and a FPR of 0.19. This should be considered a very solid performance, taking into account that it is classifying the stimulus by only taking into account the spike counts on 36 TDEs, without using the information encoded by the timing of the spikes, and without performing any learning method in order to obtain the optimal values for the TDE layer hyper-parameters.

In order to asses the reliability of the results, the experiment has been repeated for different keywords in the dataset, showing similar results (Appendix, Figures 17 and 18).



(a)

(b)

Figure 15: Comparison of the ROC curves for the classification with the formants and with the TDEs, when using the 2% of channels in the classification (a) and when using the 20% (b). The red dashed line corresponds to a random classifier.

5 Conclusion

In this thesis, the information processing in neuromorphic keyword spotting models is researched, in order to evaluate the introduction of the Time Difference Encoder to this models for improving its performance. The temporal features in the speech encoding that allow to identify the spoken words are explored, and how the TDEs can extract this features in order to improve the existing neuromorphic keyword spotting models. In the models tested, the mutual information measurements show that the TDEs can successfully encode temporal features present in the human speech by the spiking of certain TDEs in the TDE layer, easing the classification of the stimulus. Moreover, the study of the formants model show that a reduced population of TDEs, specific for the keyword that wants to be spotted, can be used to successfully perform the keyword spotting task.

5.1 Future Work

In order to develop a biologically realistic keyword spotting model, it is important to take into account that in real-world data, the sound volume of the speakers and the back-ground noise levels can show a high variability. The TDEs can be very sensitive to both factors, as the volume of the speaker is directly related to the spike counts in the cochlea channels, and the noisy spikes in the cochlea can trigger TDEs that should not be active. In this sense, the introduction of a filtering mechanism, such as a Winner-take-all network connected to the cochlea, or a noise suppressing mechanism based on lateral inhibition in the cochlea should be studied.

Bibliography

[1] M. Coath, S. Sheik, E. Chicca, G. Indiveri, S. Denham, and T. Wennekers, “A robust sound perception model suitable for neuromorphic implementation,” *Frontiers in Neuroscience*, vol. 7, p. 278, 2014.

[2] T. Rost, H. Ramachandran, M. Nawrot, and E. Chicca, “A neuromorphic approach to auditory pattern recognition in cricket phonotaxis,” *Circuit Theory and Design (ECCTD), 2013 European Conference on (Dresden)*, pp. 1–4, 12 2013.

[3] P. Blouw, X. Choo, E. Hunsberger, and C. Eliasmith, “Benchmarking keyword spotting efficiency on neuromorphic hardware,” pp. 1–8, 03 2019.

[4] T. Mikutta, “Improving keyword spotting on neuromorphic hardware using time difference encoder neurons, journal = Master’s thesis, MSc in Biomechatronics, University of Bielefeld,,” 11 2020.

[5] H. Markram, K. Meier, T. Lippert, S. Grillner, R. Frackowiak, S. Dehaene, A. Knoll, H. Sompolinsky, K. Verstreken, J. DeFelipe, S. Grant, J.-P. Changeux, and A. Saria, “Introducing the human brain project,” *Procedia Computer Science*, vol. 7, pp. 39–42, 2011. Proceedings of the 2nd European Future Technologies Conference and Exhibition 2011 (FET 11).

[6] M. B. Milde, O. J. N. Bertrand, H. Ramachandran, M. Egelhaaf, and E. Chicca, “Spiking Elementary Motion Detector in Neuromorphic Systems,” *Neural Computation*, vol. 30, pp. 2384–2417, 09 2018.

[7] G. D’Angelo, E. Janotte, T. Schoepe, J. O’Keeffe, M. B. Milde, E. Chicca, and C. Bartolozzi, “Event-based eccentric motion detection exploiting time difference encoding,” *Frontiers in Neuroscience*, vol. 14, p. 451, 2020.

[8] T. Bekolay, J. Bergstra, E. Hunsberger, T. DeWolf, T. Stewart, D. Rasmussen, X. Choo, A. Voelker, and C. Eliasmith, “Nengo: a python tool for building large-scale functional brain models,” *Frontiers in Neuroinformatics*, vol. 7, p. 48, 2014.

[9] L. Chittka and A. Brockmann, “Perception space—the final frontier,” *PLOS Biology*, vol. 3, p. null, 04 2005.

[10] M. S. A. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, “A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics,” *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2390–2412, 2009.

[11] M. Zilany, I. Bruce, and L. Carney, “Updated parameters and expanded simulation options for a model of the auditory periphery,” *The Journal of the Acoustical Society of America*, vol. 135, pp. 283–6, 01 2014.

[12] N. M. Timme and C. Lapish, “A tutorial for information theory in neuroscience,” *eNeuro*, vol. 5, no. 3, 2018.

[13] R. Ince, R. Petersen, D. Swan, and S. Panzeri, "Python for information theoretic analysis of neural data," *Frontiers in Neuroinformatics*, vol. 3, p. 4, 2009.

[14] S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen, "Correcting for the sampling bias problem in spike train information measures," *Journal of Neurophysiology*, vol. 98, no. 3, pp. 1064–1072, 2007. PMID: 17615128.

[15] I. Titze, *Principles of Voice Production*. Prentice Hall, 1994.

[16] D. P. W. Ellis, "Sinwave speech analysis/synthesis in matlab." http://www.ee.columbia.edu/ln/labrosa/matlab/sws/, 2004. Online source.

[17] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[18] "Shutterstock." https://www.shutterstock.com/image-photo/voice-controlled-smart-speaker-little-kid-1837400464, 2021. Accessed: 30.11.2021.

[19] "Wikimedia commons." https://commons.wikimedia.org/wiki/File:Neuron3.png, 2021. Accessed: 17.10.2021.

Appendix

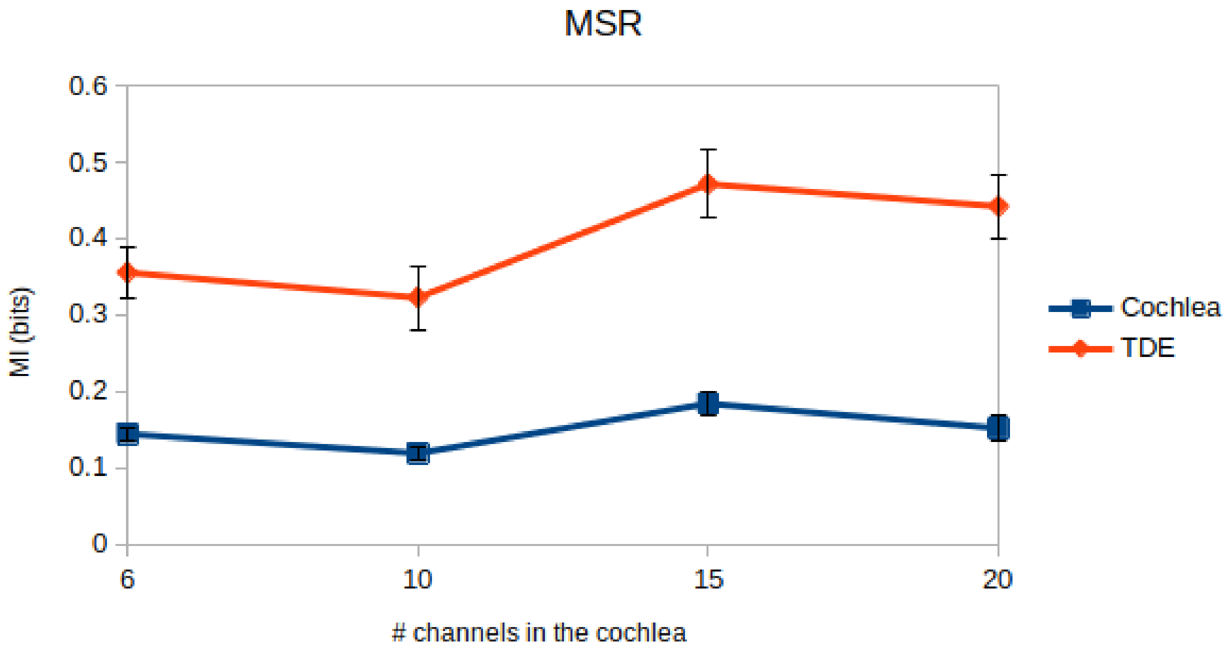
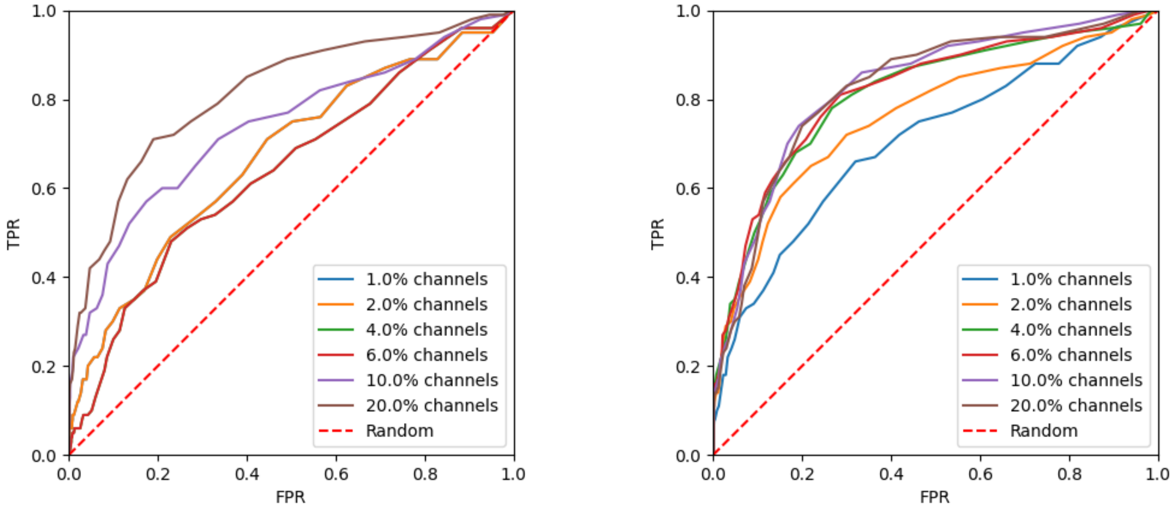
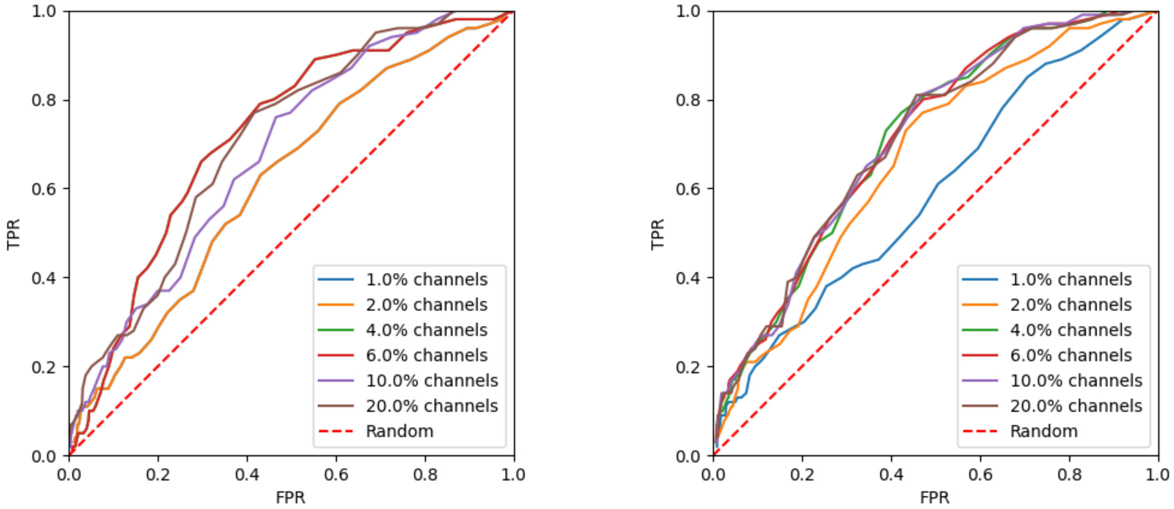


Figure 16: Mutual information between the cochlea (blue) and the TDE layer (orange) versus the stimulus presented as a function of the number of channels in the cochlea. In this case, the cochlea model contains 1 MSR fiber per channel. The errors are estimated by the standard deviation between 3 measurements with different bias correction methods



(a) (b)

Figure 17: ROC curves for the classification of the stimulus by the spike counts in the formant channels (a) and by the spike counts in the TDE layer (b) for the keyword *four*. Each curve corresponds to the percentage of the formant channels/TDEs used in the classification of the stimulus, and the red dashed line corresponds to a random classifier.



(a) (b)

Figure 18: ROC curves for the classification of the stimulus by the spike counts in the formant channels (a) and by the spike counts in the TDE layer (b) for the keyword *zero*. Each curve corresponds to the percentage of the formant channels/TDEs used in the classification of the stimulus, and the red dashed line corresponds to a random classifier.