Imperial College London

MSc Thesis

Bio-inspired Encoders for Temporal Convolutional Networks performing Speech Enhancement and Separation

Jean-Marie Lemercier

Supervised by Dr. Dan F. M. Goodman in collaboration with Dr.-Ing. Timo Gerkmann

May 2020

Imperial College: Department of Electrical and Electronic Engineering

Contents

Acknowledgements

I would like to show my deep appreciation to Dr Dan F.M. Goodman and Dr.-Ing. Timo Gerkmann for co-supervising this MSc thesis work: many thanks to Dan for his rich ideas on how to introduce biological insights and models in Artificial Intelligence methods; and to Timo for his immensely valuable inputs on the Signal Processing, Neural Networks and Auditory aspects related to this work. I am very thankful for the diverse talks we were able to have, as well as their flexible yet stimulating approach to this work as supervisors.

I also want to demonstrate my gratitude to the Department of Informatics of the University of Hamburg, and especially the Signal Processing group led by Dr.-Ing. Timo Gerkmann, for their material assistance at the benefit of my research project, and also for the numerous inputs and feedback I was given when presenting my work to the team.

I am also deeply grateful to my family for their support, and my Imperial College fellows and housemates for the enriching cross-domain chats we had regarding our respective research topics.

Finally, many thanks to Prof. Patrick Naylor for marking this work, for his project-defining advices introducing me to the Auditory field, and his remarkable Speech Processing lectures.

Introduction

Human evolution has shaped and optimized the auditory system to analyse natural sounds and especially speech. Studies have led to interesting discoveries about the early stages of the auditory processing, highlighting integration mechanisms explaining the impressive human ability to perform tasks such as Source Separation (the famous 'Cocktail-Party effect') or Speech Enhancement (removal of interfering noise and reverberant components).

For hearing-impaired listeners, those tasks are much harder to operate, explaining why it is of high interest for research to understand, model and find solutions to reproduce the operations found in normal listeners. Recently, Deep Neural Networks have been introduced in Speech Processing fields and have shown promising results regarding tasks as Speech Separation and Enhancement, by simply training on speech data, without further assumptions regarding speech structure for most of those solutions. One question naturally arises: do Neural Networks optimize their speech analysis in the same way as humans have done through evolution?

A recent example [1, 2] has shown that, when given the opportunity of learning a specific filterbank for speech analysis, a Neural Network converged to a solution which is strikingly similar to the best model we have of the human cochlea, that is, the Gammatone filterbank [3]. One practical interest of that discovery is notably that we can then freeze the filters in the Neural Network encoder to reduce the number of free parameters, enabling easier training without sacrificing optimal solutions.

Motivated by this result, we investigate another mechanism of the auditory system which is phase-locking, that is, the ability of some neurons to synchronize their firing activity to the stimulus frequency and accordingly deliver a time-coded representation of the given stimulus. We propose an encoder model enabling phase-locking to be learnt in a Neural Network, and look at whether the resulting network is using this model to learn a behaviour similar to phase-locking, or to derive another optimal strategy.

In a further step of designing optimal encoders for Neural Networks performing on speech data, we add phase-shift parameters in the analysis filterbank and train the network to learn those parameters. We show that the network converges to a solution where the maximal discriminative power is obtained in order to produce the most diverse representations out of redundant information, justifying the intuition in [2].

The remaining of this document is organized as follows: in the first section we are presenting introductory concepts of human auditory processing.

The second section is dedicated to background studies focusing on the one hand on models of the early stages of auditory processing, in particular cochlear filtering and time-coding, and on the other hand on Neural Networks dedicated to Speech Processing and especially the recent Temporal Convolutional Networks (TCN).

We will then present in a third section the first experiment of this study, introducing a model for time-coding in a TCN encoder training on Speech Separation and Speech Enhancement.

We will discuss the results of this experiment in relation with our comparison between Neural Networks and human auditory system.

In the fourth section, we will present the second contribution of this work, which focuses on extending the Gammatone filterbank used in the encoder of a TCN with a multiphase-shift module where phase factors are learnt over training. We will put the observations of this experiment in relation with results of [2].

We will briefly introduce in a fifth section future projects, introducing a model for correlation variability in Temporal Convolutional Networks.

1 Speech Processing tasks and Human auditory system description

1.1 Introduction to the Human auditory system

We briefly introduce here the major mechanisms involved in sound processing by the human auditory system, as well as the associated organs, following the order of the processing chain. We will later (see 2.1) introduce popular and novel computational models for particular subsystems of the described auditory ensemble.

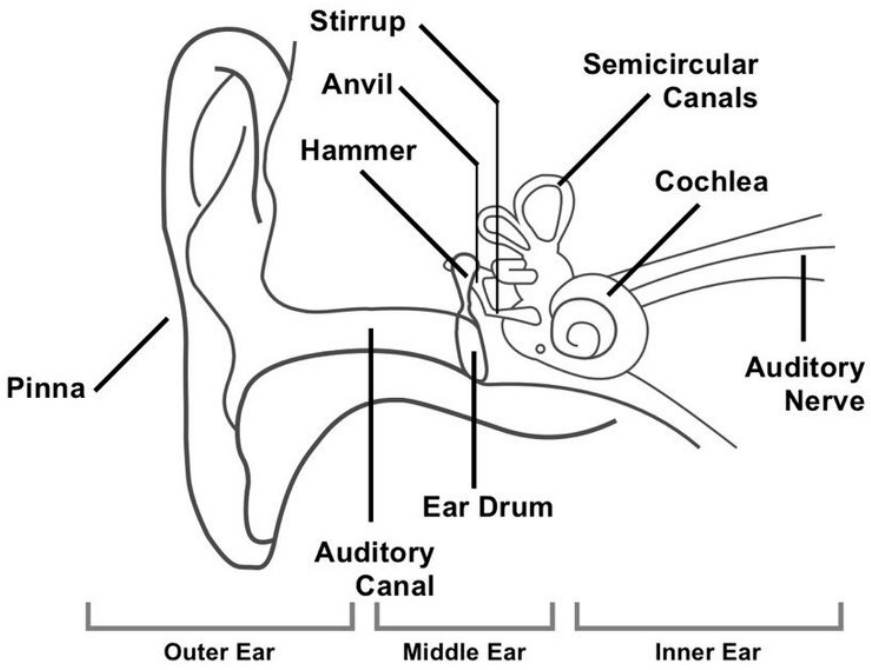


Figure 1.1: Simplified schematics of the human auditory system. From [4]

1.1.1 External and Middle Ear

The first organs exposed to sound waves inform the external ear, comprising the *pinna*, which channels the sound wave down into the external *auditory canal*. The reflections of the sound wave on the cartilage walls of the pinna produce spatial cues depending on the listener's orientation with respect to the sound source. The redirected sound wave travels through the auditory canal down to the *tympanus*, which separates the external ear from the middle ear.

The middle ear transduces the air compression wave to a solid compression wave via the *ossicular chain*, attached at its external end to the tympanic membrane. The other end is attached to the *oval window*, where the inner ear begins and where the sound wave is again transferred into a different medium, this time a fluid filling the cochlea. This ossicular chain acts as a vibration amplifier and an impedance matching device (and can also acts as a damper when a stiffening

reflex is triggered by too loud noises, in order to protect the inner ear from eventual damage)

1.1.2 Inner Ear: from the Cochlea to the Auditory Nerve

The inner ear is mainly centred around the cochlea, where the fluid-borne wave is transformed into electrical pulses and transmitted to the auditory nerve fibers.

The compression wave at the oval window becomes a travelling window along the *basilar membrane*, on the surface of which are planted the *hair cells*. The basilar membrane has particular mechanical properties which vary monotonically with respect to its length: the local dampening, width and stiffness define a place-coding scheme, where low-frequency waves travel further along the membrane than high-frequency waves. That is, we can associate a best-frequency or *Center Frequency* - which is simply the resonance frequency or the local equivalent mass-spring-damper system - to an individual hair cell, motivating the use of bandpass filterbanks (see after in 2.1.1) to model the cochlea.

The numerous hair cells (around 3500 inner hair cells and 12000 outer hair cells) have linear emerging structures called *stereocilia* which react to the stretching caused by the travelling wave by opening ionic channels. The cations entering the channels depolarize the hair cell which results in creating a receptor potential, in turn releasing neurotransmitters which finally create electrical impulses in the stem of the auditory nerve. If the inner hair cells are the "source" of the nerve signal because they are directly connected to the auditory nerve, outer hair cells are nonetheless paramount as they act as nonlinear amplifiers, which highly enhance the auditory sensible threshold by directly providing synchronous feedback to the basilar membrane in which they are implanted.

Parallel processing is achieved in another cavity structure called the *vestibular system*, instrumental in particular in producing information regarding balance and geometrical orientation of the body with respect to altitude.

1.1.3 Cortical cells involved in auditory functions

Once the signal has been converted to an electrical impulse, it travels down auditory nerve fibers, which have a *tonotopic* structure, in the sense that fibers close in space carry frequency-related content (which is directly related to the geometrical organization of the cochlea and the mechanical properties of the basilar membrane) [5, 6].

A large part of the sound cues are then processed in the Amplitude Modulation domain, especially in high-frequency ranges where the harmonic structure of the processed sounds is assumed not to be inspected, the focus being rather made on the energy envelope of the signals. Indeed, structures like impulses, onsets or offsets are often met in natural sounds and are revealing of events that might be of interest for the survival of the individual (detection of impacts, falling objects...). Some neural structures like the Inferior Colliculus, located in the brainstem, have been found to play the role of an AM-frequency filterbank, each of those cells being tuned to a certain AM Best Frequency. After that particular processing, neurons are shown to be organized in a *periodotopic* way [7], which means the signals conveyed in nerve fibers close in space are the ones processing close AM frequencies.

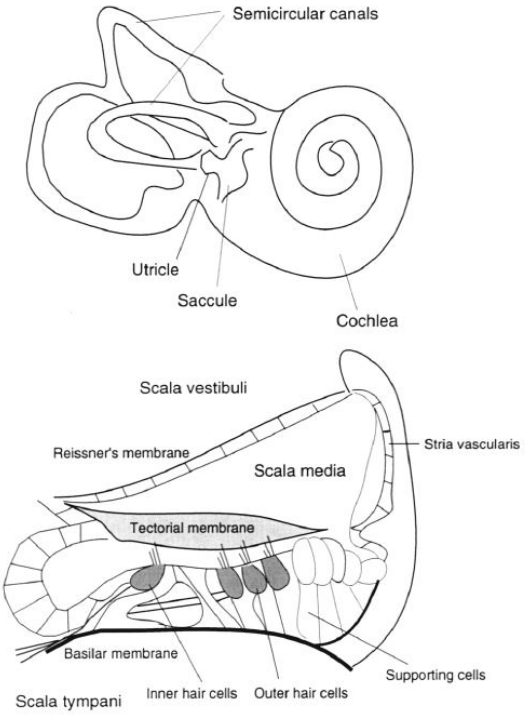


Figure 1.2: Diagram showing the innear ear structure: (top) ensemble view, (right) detail of the cochlear micro-anatomy: cross-section through the organ of Corti, encompassing the hair cells and direct neighbouring cells

After having extracted those frequency and modulation cues in each channel (left and right), signals are combined to inspect binaural correlation. It is in the brainstem and higher cortical regions that most of the spatial processing is done, such as computation of Interaural Time Differences and Interaural Intensity Differences, which are the main spatial cues extracted from binaural processing. Early representations are also provided by the vestibular system in the inner ear, and the Cochlear Nucleus in the brainstem: those representations include for instance elevation information, which is computed in the Pyramidal cells between the Cochlear Nucleus and the Inferior Colliculus. There is also evidence of a *spatiotopic* organization in the cortical cells involved in spatial sound processing [6], with spatiotopy in that case relating to structures where sounds originated from sources perceived close in space are conveyed in neighbouring neurons.

This triple organization (both tonotopic, periodotopic and spatiotopic) highlights the high efficiency of the human brain, which is then able to access and process information that is likely to be correlated in an efficient way, by making this information quickly accessible and without risking to lose or attenuate the signal during unnecessary travel in nerves.

Remark 1.1. *This concept can be related to the coalescence principle used in some algorithms training on Graphical Processing Units like [8], where the Medium Access is optimized by sorting synapse models with respect to their conduction delays and accordingly store their data contiguously. Indeed, accessing multiple data points, with a buffer size superior to the size of an individual data point, can be inefficient if you access those data points one by one, loading irrel-*

evant additional data which was contiguous to the data point of interest. However, when sorting the synapses by their conduction delays, the buffer accesses multiple data points by encompassing multiple data of interest within the same buffer (the data to be accessed simultaneously in this particular model is the discharge signals of neurons arriving at a the same time, so with similar delays).

Finally, the perceptual interpretation of all the information which was integrated in the midbrain and brain stem regions is performed in the cortical regions, where higher-level operations like source localization, speaker identification or emotion recognition are carried.

1.2 Challenging Speech Processing tasks

We introduce here three Speech Processing tasks particularly challenging for hearing impaired individuals, especially in adverse conditions, which in comparison shows the remarkable ability of a normal listener's auditory system to perform such difficult tasks, most of the time simultaneously. We introduce some mechanisms used by the auditory system to realize these tasks.

1.2.1 Speech Enhancement

Speech Enhancement or more precisely Speech Denoising describes how the auditory system discriminates a target speech signal from noise, which can range from the simplest white noise to the most unstationary noise like a drum part for instance. As we will point it out later, Speech Enhancement performance depends a lot on the input Signal-to-Noise ratio (SNR), but the nature of noise is also highly influential in how difficult the Speech Enhancement task is for an individual. The more variations and correlation the signal bears, the more difficult it is for conventional systems to separate it from target speech, and also the most disturbing it can be for the perceived audio quality of a signal. For instance, 'musical noise' (White noise with moderate time-variation of the Power Spectral Density) at a high SNR is usually much more disturbing than low-SNR white noise. Other conditions come into play such as the presence of room reverberation or the existence of a visual contact with the target speaker (audio-visual correlations play an important role in improving identification of a audio signal like speech).

If the additive noise is defined as the time-varying signal $n(t)$, the target speech $s(t)$, then the objective of Speech Enhancement is to extract $s(t)$ from the mixture $x(t) = s(t) + n(t)$.

1.2.2 Speech Separation

Speech Separation denotes the task of discriminating between multiple speakers speaking at the same time, and is more commonly designed as the 'cocktail-party' effect. A target speaker is generally identified and the other speakers considered as interferences or noise and the resulting speech is a mixture of this target and undesired speech at a given SNR. Speech Enhancement could be seen as a special case of Speech Separation where the numbers of speakers is only 2, and where the interference signal usually carries less correlation than a classical speech signal.

In favorable conditions (reasonably high SNR, anechoic environment, visual contact with target speaker), a normal listener is perfectly able to extract the statistics of the target speaker out of the noisy mixture, and then proceeds to converse with their identified interlocutor. In adverse conditions however (negative SNR, reverberant environment, visual interferences), this task becomes more challenging but normal listener's robustness to these conditions is quite impressive.

On the other hand, for hearing-impaired listeners, even in favorable conditions, this task is highly challenging: indeed, discriminating speech from another speech signal requires computing accurate statistics for both the signals, which are mixed in similar frequency regions, and both have important short-time correlation. It is shown that for instance, the ability to discriminate between two speakers of the same gender is more challenging than for two speakers of distinct gender, mainly because an important part of the discriminative power is founded on the analysis of fundamental frequency [9], and the median fundamental frequency is usually gender specific (around $100-150$ Hz for male speakers and $180-250$ Hz for female speakers). When conditions becomes more adverse, for instance in the presence of noise or reverberation, this task becomes virtually impossible for even lightly-impaired individuals.

In an anechoic environment, if M speakers are present, and the target speech signal is arbitrarily picked as $s_1(t)$, with some potential additive noise $n(t)$, then the objective of Speech Separation is to extract $s_1(t)$ from the mixture defined as $x(t) = \sum_{i=1}^{M} s_i(t) + n(t)$.

1.2.3 Speech Dereverberation

Speech Dereverberation handles the non-ideal case where the environment is not anechoic and therefore the speech signal is processed through the room acoustics, whose effect is usually modelled a linear filter characterized by its Room Impulse Response (RIR). This RIR is of course dependent on room geometry and materials, but also on the relative distance and positions of the listener to the speakers. Effects of reverberation are often separated into two parts: early reflections and late reverberation (when all the reflections accumulate and merge into some diffuse field). Speech Dereverberation can be seen as a special case of Speech Enhancement where the noise is not additive but highly correlated to the speaker. For normal listeners, reverberant acoustics are not too challenging (early reflections can even sometimes increase speech intelligibility of the speech, because of the gain in volume). For hearing-impaired listeners however, it dramatically impacts all efforts of speech recognition, especially when other adverse conditions come into play (additive noise, multiple speakers...).

If the RIR defined at the listener's position is the sequence of taps $\{h_1, \dots, h_{K_s}\}$, the target speech $s(t)$, with potential additive background noise $n(t)$, then the objective of Single-Channel Speech Dereverberation is to extract $s(t)$ from the mixture $x(t) = \sum_{\tau=1}^{K} h_\tau s(t-\tau) + n(t)$.

Remark 1.2. *We will not focus on Speech Dereverberation in this study.*

2 Background and related work

2.1 Modelling early stages of auditory processing

2.1.1 Place-coding and Cochlear filtering

Observations of the human cochlea lead to considering the existence of a place-coding scheme related to the composition of the basilar membrane: we mentioned in 1.1.2 that this organization results from the existence of a frequency-tuning characteristic of the sensors in the organ of Corti, namely the inner hair cells. The models used are therefore aimed at extracting narrowband content in separated channels, and thus are generally bandpass filterbanks. Several subjective studies have shown that there is a nonlinear warping between the classical Fourier frequency domain and the human perceptual frequency scale [10, 3], which encourages to define a new frequency scale where Center Frequencies will be uniformly distributed. The choice of this scale as well as the actual shape of each individual filter defines the type of filterbank that will be used for analysis, and eventually synthesis depending on the application.

Fourier Filterbank

A classical filterbank used in audio applications is the Fourier filterbank, where Center Frequencies are uniformly distributed along the Fourier scale and filters all have the same bandwidth and can partially overlap. Filter extract the Fourier transform of the short-time windowed signal (classical window functions include Hanning, Hamming, Rectangular functions...). An example of a 32-channel Fourier filterbank is given on fig.2.1a.

Mel-scale Filterbank

More widely used in audio applications is the Mel-scale triangular filterbank, which uses a non-linear warping of the frequency scale called the Mel-scale, based on a psychoacoustical test by [10]. The most common filter shape is triangular in the Mel domain, (but some studies also use a Gaussian filter shape [11]) and make the filters overlap in the frequency-domain, trading some discriminative power for a regularity of the analysis (it is shown in [12] that there is some evidence for *frequency leakage* between neighbouring auditory nerve fibers, which means that the hair cells themselves are not perfectly selective but their analysis frequency bands overlap between each other). The warping relation is defined in eq.2.1: center-frequencies are considered uniform on that Mel-scale. The filter envelope in the Fourier domain is defined in 2.2 (We added a factor $\frac{1}{b_k}$ for conserving the RMS power constant across filters, since the bandwidth b_k is also defined on the Mel-scale, so it is increasing with frequency).

$$CF^{(Mel)} = 2595\ \log_{10}(1 + \frac{CF^{(Hz)}}{700}) \tag{2.1}$$

$$H_k(f) = \begin{cases} k \text{ even:} & \begin{cases} \frac{1}{b_k}\left[\frac{2}{b_k}(f - CF_k) + 1\right] \\ \frac{1}{b_k}\left[\frac{2}{b_k}(CF_k - f) + 1\right] \\ 0 \end{cases} \\ k \text{ odd:} & \begin{cases} \frac{1}{b_k}\left[\frac{2}{b_k}(f - CF_{k-1})\right] \\ \frac{1}{b_k}\left[\frac{2}{b_k}(CF_k - f)\right] \\ 0 \end{cases} \end{cases}$$

An example of a 32-channel Mel-scale triangular filterbank frequency-representation is given on fig.2.1b.

Gammatone Filterbank

Later on, the Gammatone filterbank has proved efficient in representing the frequency decomposition performed by the cochlea [3, 13], with its output being qualified as *cochleagram*, whose surface (as a function of Center Frequency and time) is supposed to model the motion of the basilar membrane. The impulse response of the Gammatone filterbank is supposed to provide physiologists with an expression of the reverse correlation between the firing of an auditory nerve fiber and the input waveform [3, 14], (which takes into account the observed physiological data more than a triangular or Gaussian envelope as in the previous filterbanks). The response is given for a certain Center Frequency by a gamma function multiplied by a cosine tone (see eq.2.2). The resulting filterbank is a linear bandpass filterbank with non-linear warping of the frequency axis.

$$h_k(t) = a_k t^{n-1} cos(2\pi CF_k^{(ERB)} t + \phi) e^{-2\pi b_k t} \tag{2.2}$$

where in eq.2.2 are:

- n the order of the filter (determined between 3 and 5 in the human [14], but is to be lowered to 2 in our implementation by need of a short filter impulse response)
- a_k is a scaling factor aimed at normalizing the RMS power across filters
- $CF_k^{(ERB)}$ the Center Frequency on the Equivalent Rectangle Bandwidth scale (see thereafter)
- b_k an approximated bandwidth given proportional to the Equivalent Rectangle Bandwidth (ERB)

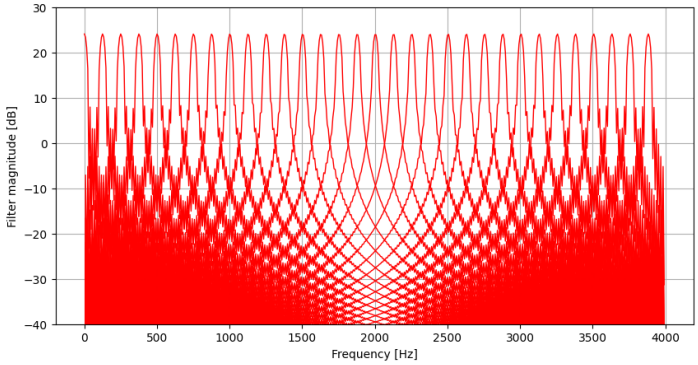
The equivalent rectangle bandwidth is issued from an approximation modelling the cochlear filterbank by an ideal bandpass linear filterbank. In [15], an affine approximation fitting human empirical data links a CF to a filter with bandwidth $ERB(CF)$ by the following affine relation:

$$ERB(CF^{(Hz)}) = 24.7\ (1 + 0.00437\ CF^{(Hz)}) \tag{2.3}$$

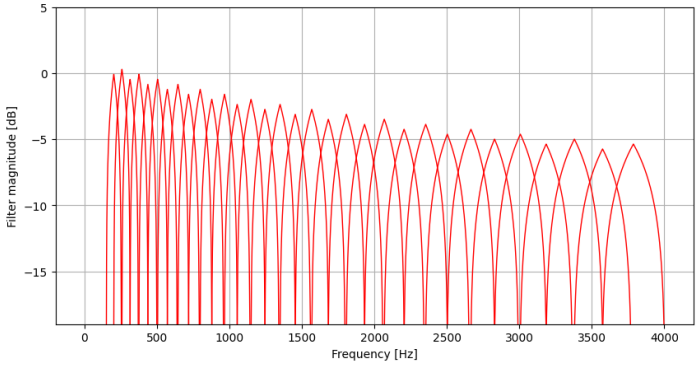
In the Gammatone filterbank, the CF distribution is considered uniform on the ERB scale following the previous ideal-rectangular-bandpass-filterbank assumption. This links the ERB scale to the Hertz scale by a logarithmic relation (from integrating the inverse of eq.2.3 with respect to the CF, to get the rate of change from an ERB Center Frequency equivalent to the next one) [15]:

$$CF^{(ERB)} = 9.26\ log_{10}(1 + 0.00437\ CF^{(Hz)}) \tag{2.4}$$

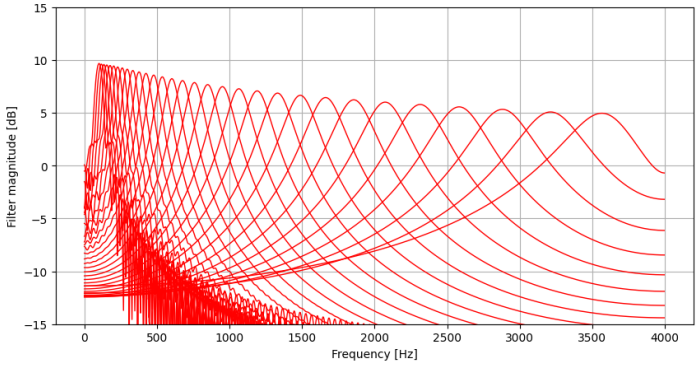
An example of a 32-channel Gammatone filterbank frequency-representation is given on fig.2.1c



(a) 32-channel STFT filterbank frequency response.



(b) 32-channel Mel-scale filterbank frequency response.



(c) 32-channel Gammatone filterbank frequency response

Figure 2.1: Frequency responses of some filterbanks commonly used in audio-processing. Sampling frequency: $8000Hz$

The use of this sophisticated, bio-inspired filterbank in auditory modelling and processing has shown some great results: for instance recently, [2] have shown that the filterbank learnt from a Conv-TasNet [1] linear encoder performing single-channel Speech Separation is very close to the Gammatone filterbank. they showed improved performances when replacing this auto-encoder by a fixed (Multi-phase) Gammatone filterbank, which is an approach we consider in section 4.

2.1.2 Time-coding and Phase-Locking

A definition of Phase-Locking

The auditory system uses many different mechanisms for the same Speech Processing tasks, which is assumed to be one of the reasons for the impressive robustness of hearing abilities in adverse conditions, because of the ability to integrate redundant information with diverse processes. For instance, monaural processing such as pitch perception is believed to use both the previous place-coding mechanism and some time-coding scheme, which is implemented by a *phase-locking* behaviour observed in many animal species [16, 17, 18]. In addition, binaural processing tasks such as Interaural Time Difference computations have been shown to rely heavily on phase-locking.

Phase-locking can be described as following: for stimulus tones, observed Auditory Nerve neurons output a discharge during only one half-wave of the tone, and decay during the second half: this phenomenon can be modelled by a Rectified Linear Unit (ReLU) as a simple half-wave rectification operation. Thanks to that mechanism, neurons in the auditory nerve and beyond are then able to represent variations of the signals (like stimulus fundamental frequency, or interaural time difference resulting in modulation frequencies) as a function of time, thus creating a "time-coding" scheme.

Fig.2.2, (bottom) gives the discharge rate of a population of 50 neurons in the Auditory Nerve following a tone input (represented in the middle part of the figure), thus giving an early representation of the signals in the auditory system,. The neuron response is given as the histogram of spikes globally fired at the neurons output (the spike repetition is given in the top part of the figure: one dot on of coordinates (t, k) represents a spike fired by the neuron k at the time t), and reflects a structure which follows the positive envelope of the signal, motivating the previous idea of using a ReLU in the pre-processing unit to model phase-locking.

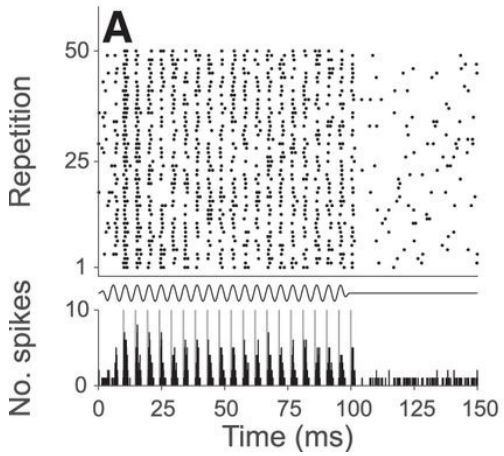


Figure 2.2: Single unit recordings displaying Phase-locking for low-frequency tones

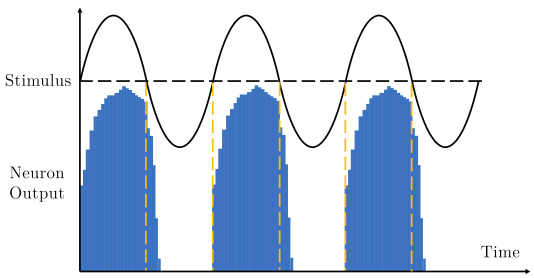


Figure 2.3: Schematics of phase-locking to a low-frequency stimulus

Phase-locked intensity can be measured by a coefficient of synchronization [16] which is simply the fraction of number of spikes during the first half period to the total number of spikes during a full stimulus period:

$$C_s = \frac{\sum_{t\in[0,\frac{T}{2}]} \mathbf{1}_{spike}(t)}{\sum_{t\in[0,T]} \mathbf{1}_{spike}(t)} \in [0.5-\epsilon, 1] \tag{2.5}$$

However, Vector Strength is more widely used in recent literature as [17, 18], and is defined by the following relation in [19]:

Definition 2.1. *Vector Strength*

For some events defined at instants $\{t_n\}_{0\leq n\leq N-1}$ *(in our case, spikes fired by neural cells), we define the* ***Vector Strength*** *to be the measure of synchronization between the vector of these time events and a periodic stimulus of fundamental frequency* f_0*:*

$$r(f_0) = |\frac{1}{N}\sum_{n=0}^{N-1} e^{j2\pi f_0 t_n}| \tag{2.6}$$

If the events are perfectly synchronized to the stimulus frequency and consequently gathered in the first semi-period, the vector strength will be maximal (closer to 1), and if the events are closer to an activity spreaded on a whole period $[\frac{k}{f_0}, \frac{k+1}{f_0}]$*, the Vector Strength will be closer to 0.*

Phase-locking in animals

Studies listed in [18, 17] show that the Vector Strength decreases with frequency, which is often modelled by a lowpass fashion, with order and DC gain approximately constant across species, and a species-dependent cutoff frequency [20]. This cutoff frequency has been estimated to about $4.5kHz$ for the squirrel monkey [16], $2kHz$ in the chinchilla [21] and around the same for the cat [22]. Species are also directly compared in [23], see fig.2.4

To account for this phenomenon, [16] states that the refractoriness of the AN fiber (influencing the decay of the discharge response) does not intervene in the discharge rate of the response, but limits the phase-locking phenomenon to the lower frequencies (which is understandable since the fiber has not discharged its potential that already a new period of the stimulus is causing a discharge). The authors also claim that the probability of spike firing is an increasing, power-like function of the stimulus strength and a linearly decreasing function of the stimulus frequency. Finally, they make the statement that this phase-locking phenomenon takes place in the periphery of the AN fiber.

Phase-locking in the human

Phase-locking has been shown to be used by the human brain for binaural processing up to $1.5kHz$ typically for Interaural Time Difference computation [25], and possibly for monaural processing. However, there is still some debate over whether this upper limit is effectively *(i)*- that of pure phase-locking ability of human neurons, that is, how efficient are human cells at analysing Time Fine Structure (TFS) up to that frequency; or *(ii)*- the frequency above which the known mechanisms do not rely on a time-coding scheme anymore, but rather use another mechanism, like place-coding [18].

[18] compiles different points of view on the subject of frequency-dependent TFS analysis abilities in the human. It mentions in particular experiments from [26] on measuring Different Limens for Frequency (a subjective measure of pitch variation detection by the analysis of high harmonics of a stimulus). It is shown there that the DLF is increasing with frequency up to

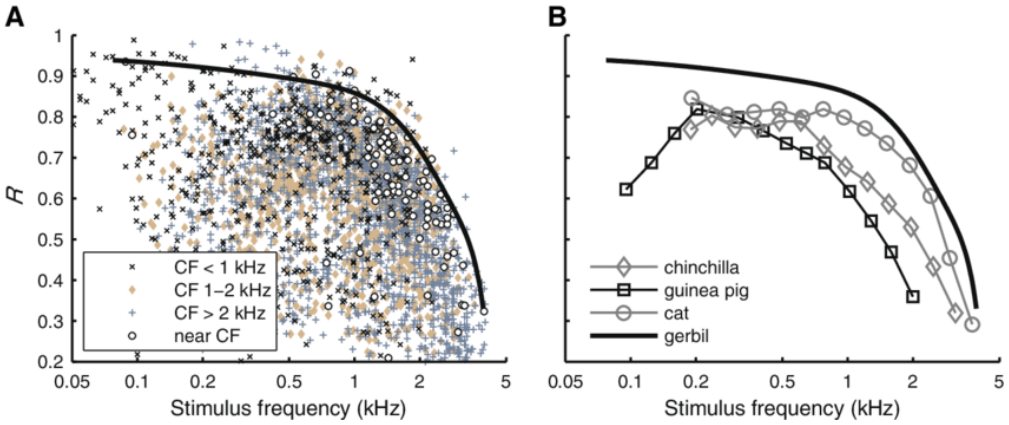


Figure 2.4: Vector Strength in function of Stimulus Frequency for pure tones stimuli, from [23] based on compiling studies on 4 different species [20, 23, 24]

around $8kHz$, which is expressive of a time-coding scheme [26], whereas it plateaus for higher frequencies (up to $10kHz$ which is the maximal frequency used in the study) which is considered to be characteristic of a place-coding scheme (see fig.2.5). This tends to highlight that time-coding is not the only scheme used in pitch-perception, and that place-coding would be used for higher-frequency pitch analysis whereas time-coding would be used for lower-frequency.

Oxenham & al. on the other hand, insists on the fact that neural mechanisms involving phase-locking for binaural processes have been clearly identified, whereas it is not the case for monaural processing [18], where only subjective data are available, which have them stating that the only known upper frequency limit for phase-locking is that of ITD computation, namely $1.5kHz$. They also refer to [12] to emphasize on the fact that having perceptive performances relying on several mechanisms like time- and rate-coding really does not help in identifying the actual boundaries of the said mechanisms [27].

Taking in account all of the previous statements does not help conclude over the upper limit of frequency for phase-locking ability, and one of the goals of our study is therefore to investigate whether phase-locking can be modelled and learnt by a Neural Network encoder, by training toward speech processing tasks such as Speech Separation and Speech Enhancement. Possible outcomes of this experiment are that *(i)*- the Neural Network learns through the proposed encoding some behaviour directly comparable to phase-locking in the human; *(ii)*- the Neural Network optimizes its functioning with respect to the proposed encoding in a way which is not interpretable in terms of biological meaning.

2.1.3 Further neural representations

We broadly introduce here popular mechanisms investigated in Auditory system modelling and mention classical models used to describe those mechanisms. We relate to those mechanisms in the context of the DNN-based methods, justifying the use of Neural Networks for Speech Processing.

Envelope extraction

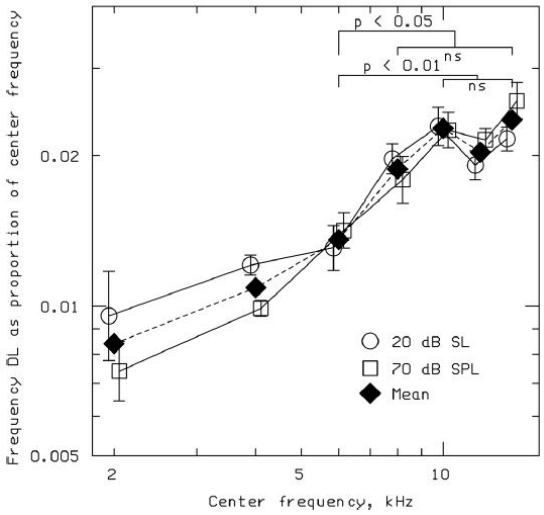


Figure 2.5: Geometric mean DLFs across nine subjects, plotted as a proportion of center frequency. From [26]

As we mentioned it in the previous subsection, Time Fine Structure is often assumed not to be used by neural mechanisms at high frequencies. This has been often linked as well to the behaviour observed in the transduction of mechanical travelling waves along the basilar membrane into receptor potential in the inner hair cells [28, 29] and is generally modelled as follows:

Considering the waveform S_k found in a single frequency-subband k (after cochlear filtering), the waveform is amplified by a linear gain in dB G_{dB}, passes through half-wave-rectification, and then is compressed by a power-like function of exponent γ (2.7). Finally, the resulting positive waveform is passed through a first-order lowpass-filter (2.8), of cutoff-frequency equal to $\frac{1}{\tau} = 1kHz$ in [28, 5], which results in an non-linear envelope extraction for high-frequency subbands, and non-linear compression and rectification for lower-frequency subbands.

$$S_k^{(comp)}(t) = \left[max\left(10^{\frac{G_{dB}}{20}} S_k(t), 0\right)\right]^{\gamma} \tag{2.7}$$

$$\tau \frac{dS_k^{(env)}}{dt}(t) = S_k^{(comp)}(t) - S_k^{(env)}(t) \tag{2.8}$$

The gain and compression operations are assumed to describe the mechanical coupling between the basilar membrane and the outer hair cells. Indeed on the one hand the amplification induced by outer hair cells feedback to the basilar membrane is non-linear, and on the other, the mechanical properties of basilar membrane induce non-linearities when transforming the compression waves into travelling waves (see 1.1.2).

The half-wave-rectification on the other hand is modelled based on the observed relation between stretching of the base of an inner hair cell stereocilium and the depolarization of the cell [5]. The signal is then integrated by an equivalent capacitance of the cell membrane, which leads to this first-order lowpass filter model. Although the real phenomenon is likely to induce more non-linearities, this approximate model usually matches experimental data well [5, 30].

In that way, it is clear to notice that the result of these operations is to phase-lock the signal at low-frequencies, and to extract the envelope energy for higher-frequencies, which is consistent with experimental data on animals [31, 22, 21] showing that high-frequency content is almost exclusively analysed in terms of Amplitude Modulation. This adds up to another possible source of phase-locking linked to the refractoriness of auditory neurons mentioned earlier [16].

In terms of DNN implementations, a Convolutional Neural Network can clearly learn those operations if these reveal to be optimal, by simply implementing *i)*- lowpass filters with 1D-convolutional layers (approximating IIR filters by FIR filters); *(ii)*- half-wave rectification with Rectified Linear Units (ReLU) ; *(iii)*- power-like compression with sigmoid non-linearities (which would have an additional saturation effect).

Adaptation Mechanism

The term adaptation refers here to the mechanism taking place at the synapse between inner hair cells and auditory nerve fibers, that is, the earliest production of electrical impulse. Usual models consider a relation between the number of neurotransmitters to be released by effect of a depolarization stimulus and the amplitude of this stimulus, more or less sophisticated (single [29], or triple [32] vesicle-pool models have been considered). More simple models like [28] simply include inverted first-order feedback loops which result approximately in computing the logarithm of the depolarization signal.

We will not got into detail of these models but simply notice that all of them involve lowpass-filtering, exponential or power-like non-linearities, which again can be accounted for and learnt in Convolutional Neural Networks.

AM analysis

Most of the human hearing focuses on the $2kHz-4kHz$ frequency-range [5] and as stated earlier, this range is assumed not to be covered by TFS analysis. Therefore, most of the subsequent processing is done on the envelope of the signal, and that is basically Amplitude Modulation analysis, a behaviour observed in Inferior Colliculi cells. Some models focus on that critical AM analysis and as in [28] propose to model it with a Modulation filterbank, which follows the variations of the extracted envelope for AM frequencies going from 1 Hz to 200Hz, and then enables to obtain a full Center-Frequency / Modulation-Frequency bivariate analysis.

If we relate to this modulation filterbank in the Convolutional Neural Network framework, no module is explicitally implemented so as to perform such operation, but using *dilated* convolutional layers like what is done in Temporal Convolutional Networks (as we will present it in the next subsection 2.2) can virtually operate in a similar fashion. Indeed, using multiple dilated convolutional layers with different dilation factors is equivalent to analyzing the signal with the same convolutional subnetworks but on different timescales, which is somewhat similar to a modulation filterbank where the ratios between the AM-frequency analysis values are determined by the dilation ratios.

2.2 Single-Channel Speech Processing with Temporal Convolutional Networks

We reference here the main computational methods which have been proposed to perform automatic Speech Separation and Speech Enhancement, including conventional methods and DNN-based approaches. We will focus more precisely on Temporal Convolutional Networks, a state-of-the-art DNN-based approach using dilated convolutional layers, capturing speech statistics on multiple time scales.

2.2.1 Conventional Methods

We will introduce briefly the main conventional methods (as opposed to AI-based) used for Speech Enhancement and Spectral Subtraction, mainly in order to mention the shortcomings of these methods overcome by Neural Networks. Conventional methods mostly operate in the spectral domain, and can be model-based, which mean they assume a certain probability distribution for clean speech and/or interferences, or more generally rely on approximations regarding those signals.

Speech Enhancement

For Speech Enhancement, the main conventional paradigms historically used are Wiener Filtering, Spectral Subtraction and Statistical Estimators.

Wiener filtering is the a simple statistics-based approach maximizing the output SNR: it is often bringing many distortions to the signal when the noise is not stationary Gaussian.

Spectral Subtraction [33, 34], focuses on estimating the Noise signal Power Spectral Density (PSD), for instance when voice inactivity is detected, and then simply subtract it to the mixture PSD to obtain the clean signal spectrum.

Criterion-based estimators, on the other hand, are designed to extract the statistics of clean speech and the interfering signals, by first modelling the distributions of the signal spectral features - most of the time as Gaussian Mixtures - and then using a Maximum Likelihood or a MMSE-based [35, 36] criterion to iteratively compute the prior distribution of the clean speech to extract.

Speech Separation

Although conventional Speech Separation is usually based on Multi-Channel methods (Spatial Filtering being the most classical one), Single-Channel conventional methods also exist, including Pitch-tracking with source-filters [37], model-based approaches [38, 39], or deterministic Time-Frequency masking [40]. Pitch-tracking is still a paradigm in use since it was shown in various studies including [2] to be one of the main mechanisms used by the auditory system in speech separation, explaining why it is more difficult to separate the pitch from two persons of different gender than that of same gender (as median pitch is a gender-related characteristic).

However, the most recent conventional methods for Speech Separation are based on Non-negative Matrix Factorization (NMF) [41, 42, 43, 44] and this technique has also been used for Speech Enhancement [45]. NMF is a non-negative version of Independent Component Analysis: it aims at finding the optimal decomposition - generally under a Least Squares or Kullblack-Leibler Divergence criterion - of the mixture spectrogram as the product of two matrixes identified as the matrix of basis signals and activations. In classical NMF, the basis signals are the spectra

of the different units identified (in that example, some notes played on a piano), and activation matrix tracks the magnitude of these units through the time dimension. A variant of that approach called Non-Negative Matrix Factor Deconvolution [42] overcomes the main issue with NMF: when the identified units are not static enough (as in speech for instance, where variation across phonemes is high), a simple basis signals - activation decomposition as in NMF cannot properly represent the signal.

Even though those conventional methods can combine different improvements such as sophisticated models for noise or speech priors [46] with improved estimation algorithms [47] and phase reconstruction [48], they are usually not robust enough to high noise non-stationarity for Speech Enhancement, or to other adverse conditions like background noise in the Speech Separation task or reverberant acoustics for Separation and Enhancement.

2.2.2 AI-based approaches

As a solution for more robustness, especially to noise non-stationarity in Speech Enhancement, approaches using Neural Networks were introduced. They mainly relied on feed-forward architectures at the beginning [49, 50], but the interest rapidly shifted toward networks suited for sequence-modelling were studied, such as Recurrent Neural Networks (RNN) [51, 52, 53].

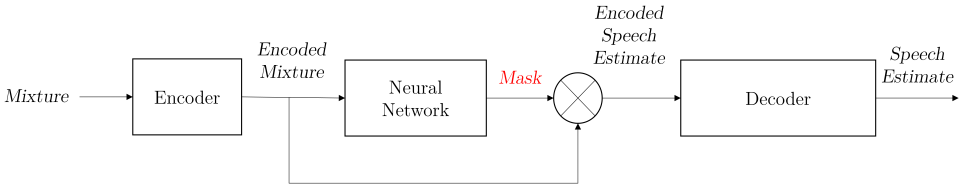
The main part of AI-based approaches focus on Time-Frequency masking and Spectral mapping as a target for training Neural Networks (see fig.2.6): both these approaches operate in the Short-Time Fourier domain and try to learn a Time-Frequency map informative of the signal and/or the interference.

In Time-Frequency masking, a Neural Network tries to learn a *mask*, that is, a map giving the energy proportion between the target signal and the interferences, at the time-frequency-bin level. This mask is then multiplied to the mixture input spectrogram to yield the target spectrogram estimate. Classical mask designs include the Ideal Binary Mask (IBM), the Ideal Ratio Mask (IRM) or the Phase-Sensitive Mask (PSM), and a complex-IRM is also defined in [50]. The two latter masks include the phase information in their design, as opposed to the two first ones, focusing only on STFT magnitude.

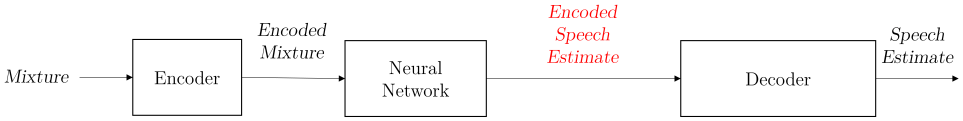
In Spectral mapping, the training target is not a mask but the Neural Network rather directly learns to map the mixture spectrogram to the target spectrogram (or any other subband-based representation such as Time-Mel-scale-frequency spectrogram or cochleagram).

Neural Networks implementing Time-Frequency masking or Spectral mapping spanned from classical feed-forward networks (Multi-Layer Perceptrons, Deep feed-forward networks...) to more recent RNN with Long-Short-Term Memory units (LSTM). These latter networks have been shown to better model sequences like speech as they explicitly take into account the temporal dependency of the signal, which is present on multiple scales for speech. Speech is indeed highly correlated at the phoneme level ($\sim 10ms$), especially for voiced speech, and long-scale variations need to be tracked when studying semantics or identifying a speaker for instance.

To include analysis of those temporal variations, a RNN proceeds as follows: given a hidden layer $h^{(k)}$ with weights and bias $(W^{(k)}, b^{(k)})$, the activation function σ of the hidden layer will take both as inputs *(i)*- the signal at the current time-frame x_t and *(ii)*- the output of the hidden layer at the previous time-frame $h^{(k)}(x_{t-1})$:



(a) Schematics of a Speech Processing using a Neural Network trained on Mask target



(b) Schematics of a Speech Processing using a Neural Network trained on Mapping target

Figure 2.6: Comparison of Time-Frequency Masking and Spectral mapping approaches for DNN-based Speech Processing solution. Output of the separating Neural Network is highlighted in red

$$h^{(k)}(x_t) = \sigma(W^{(k)}x_t + b^{(k)} + U^{(k)}h^{(k)}(x_{t-1})) \tag{2.9}$$

This can be seen as a time-varying layer with parameters $(W^{(k)}, b^{(k)}, U^{(k)})$ but the parameters are actually all stored for each time-step and gradients are sequentially dependent from one time-step to another by eq.2.9, so the training procedure needs to unfold the different layers in order to update the parameters. The advantage of this is that time-correlation is explicitly encoded, but it is computationally expensive to train these networks as the number of parameters gets rapidly large if you consider multiple layers and numerous time-steps.

Improvements have been brought to this classical RNN architecture with the use of LSTM units, and bidirectional structure [54, 52].

LSTM units use multiple versions of the inputs which pass through different gates: an *input*, *output* and *forget* gate, which enables to do a selection at the layer-level of the meaningful information (it can be seen as applying a local mask to the input).

Bidirectional structure for RNN has been introduced to include a time-reverse sequence-modelling dimension: to each recurrent layer performing time-forward sequence-modelling is added a similar layer taking the time-reverse sequence as its input. This enables the network to track non-causal effects, which can be very useful given the short-time and long-time correlations of speech.

Finally, state of the art performances for Speech Separation were achieved when combining those Recurrent Neural Networks with an unsupervised approach called Deep Clustering [55] (and its variant Deep Attractor Network [56]). Deep Clustering consists in finding an embedding

for the mixture where the sources can be easily discriminated at evaluation time via a simple k-mean clustering algorithm. Before the introduction of Temporal Convolutional Networks, the state of the art for Speech Separation was the network Chimera++ [57] which consists in a Bidirectional-LSTM-based architecture where a branching is operated. One of the branches learns the embedding for Deep Clustering and the other learns a Time-Frequency mask for mask-based Speech Separation, the two targets being jointly trained via a hybrid loss function.

2.2.3 Temporal Convolutional Networks

As stated in the previous section, recurrent structures are useful in their ability to model correlation in a time-series, but they are hard to train, which usually limits the number of layers in a network to 2 or 3 [52, 56]. [58] states and demonstrates that Convolutional Networks can be similarly used for sequence modelling (they are called in that case Temporal Convolutional Networks (TCN)) and outperform their Recurrent counterparts on many tasks and datasets, while being much easier to train. The new state of the art performance for Speech Separation is achieved by Wavesplit [59], a network combining a TCN-based separator and Deep Clustering with Permutation-Invariant-Training.

The pioneering example of that change of paradigm toward sequence modelling in the audio domain is Google's WaveNet [60], an audio synthesis network that achieved top performances in perceived quality using only convolution-based layers. Many TCN architectures are inspired from WaveNet and are adapted to different sequence modelling tasks such as Speech Enhancement and Speech Separation. We will present the paramount components of a TCN, and will focus on Conv-TasNet [1], which set the path for many state-of-the-art TCN approaches to Speech Separation.

TCN Architecture

A TCN is a Neural Network with the following characteristics or components:

- It is strictly composed of **Convolutional layers**: no recurrent or full-connected layer is used (the latter is replaced by a pointwise convolutional layer). Convolutions with different kernel sizes can be used within the same network, and these convolutions can be causal [1, 61] or non-causal [60], depending on the application (whether it must provide real-time ability or not).

- **Dilated Convolutions** are used: this is perhaps the most important component accounting for TCN sequence-modelling performances, as they indicate a local time-scale for analysis by the network. Dilated Convolution is denoted as $*_d$ and operates as follows: the input $\mathbf{x} \in \mathbb{R}^T$ is convolved with a filter $\mathbf{h} \in \mathbb{R}^L$ taking into account a dilation factor d which specifies the size of the shift-step during the convolution:

$$(\mathbf{x} *_d \mathbf{h})_t = \sum_{\tau=0}^{L-1} h_\tau x_{t-d\times\tau} \tag{2.10}$$

  With this operation, it is possible to choose the size of the *receptive field* of the convolutional layer (that is, the time window processed by the dilated filter at the current sample), which has the double advantage of potentially *(i)*- having a virtually large receptive field without having to increase the size of the filter ; *(ii)*- perform a multi-scale analysis, by

building layers with different dilation factors.

- **Gated activations** can be present [60] similarly to LSTM: those activations apply a local mask to the output of a block, that mask being itself an activated version (Sigmoid, Tanh...) of the block input (or the output of one of the block's sublayers, called the *gate*).

If we consider a layer with input $\mathbf{x}$, a main function h (in our typical case a convolution) with corresponding main activation σ_h ; and a gate function g with corresponding activation σ_g, the output $\mathbf{z}$ of the gated activation will be:

$$\mathbf{z} = \sigma_h(h(\mathbf{x})) \odot \sigma_g(g(\mathbf{x})) \qquad (2.11)$$

with $\odot$ denoting the Hadamard (pointwise) product.

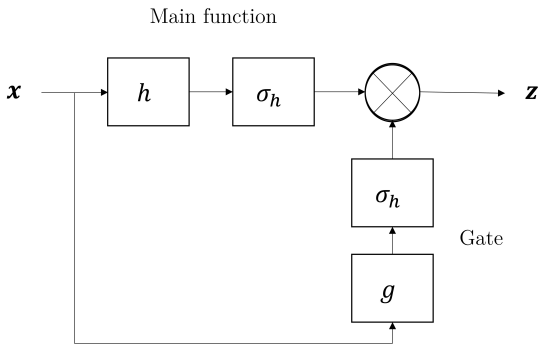


Figure 2.7: Generic schematics of a gating activated block

In the case of LSTM, it has been shown that this gating mechanism enables to get rid of the vanishing gradient issue in RNN, which make them easier to train than basic RNNs. In the context of TCN however, there is no directly identified advantage to this mechanism, other than offering another possibility for the network to select the output information it wants to forward with regard to the initial context of the considered layer.

- **Residual connections** are also one of the major component in TCN. A residual block, using residual (or skip) connections, adds the ouput of its layers to the input before the activation function, meaning the learnt transformation is actually the difference of the layer operations to identity mapping (called the residue), rather than the layer operations themselves. This has been shown to facilitate training in very Deep networks [62] by addressing the degradation problem, which translates to degrading test *and* training error as the number of layers increases.

The corresponding operation performed in residual block is then, with h being the main layers operation, σ the activation function and $\mathbf{x}$ and $\mathbf{z}$ the input and output respectively:

$$\mathbf{z} = \sigma(\mathbf{x} + h(\mathbf{x})) \qquad (2.12)$$

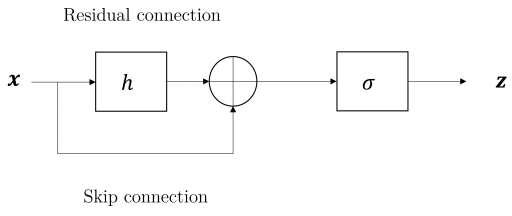
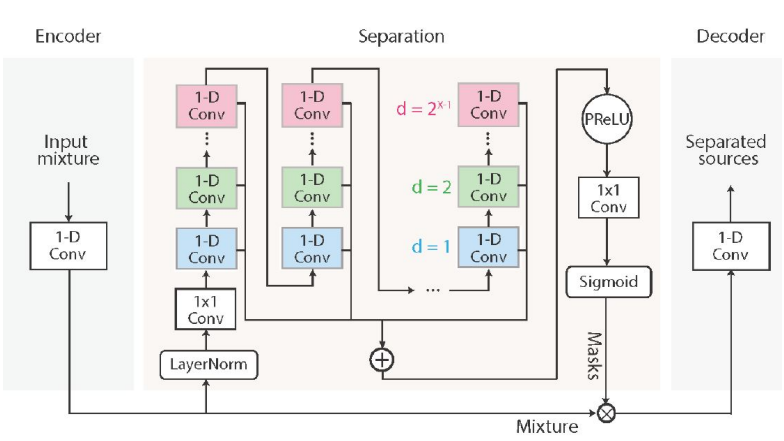


Figure 2.8: Generic schematics of a residual block

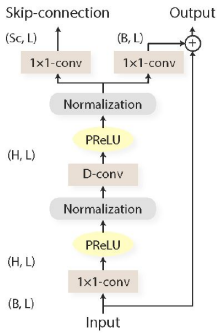
Conv-TasNet

Conv-TasNet [1] is an improved version of TasNet [63]: like its predecessor it is a Speech Separation algorithm training on masks (see fig.2.6a) but it combines an the convolutional encoder/decoder structure with a TCN composed of R stacks of X dilation convolutional residual blocks called "1-D Conv" (see fig.2.9), instead of a LSTM network as in TasNet.

In each of these blocks, the input is passed through pointwise convolution (denoted as 1×1 in the figure), and dilated depthwise separable convolution (denoted as D-conv) having a dilation factor of 2^i, $i \in 0, \ldots, X-1$. Then the output is separated in 2 branches: *(i)*- a residual structure where the layer output is fed to a pointwise convolution, summed to the input and fed to the next block; *(ii)*- a skip connection where the layer output is fed to a pointwise 1D-convolution and will directly be summed to all other skip connections to form the total output of the network. These skip connections are sometimes not used, depending on the implementations, in that case the output of the total network is the residual output of the last block.



(a) Conv-TasNet architecture overview. From [1]



(b) Structure of a "1D-Conv" Block used in Conv-TasNet. From [1]

Figure 2.9: Architecture and components of Conv-TasNet as described in [1]

The encoder and decoder in Conv-TasNet are simple convolutional layers which basically learn a filterbank composed of N filters of length L with half-segment overlapping. The information is therefore distributed in N channels, and is later processed in the TCN *across* channels (the convolutions in the TCN are not *separable*, the operation can be seen as a matrix product by a weight matrix which is not diagonal but filled), which is fundamental. Indeed [28] proposed an early model for auditory processing where the information from each Gammatone filterbank channel was independently processed, using same fixed parameters for the modulation filterbank and adaptation mechanism in each channel, thus forbidding any consideration of correlation between channels and any frequency-dependency in the processing. As opposed to this, [12] showed, by proposing a model for FM and AM analysis based on physiological observations, that human experimental data was much better fitted when operations across frequency channels were allowed.

2.2.4 Role of encoders/decoders in Speech processing tasks

We focus on this study on the role of the encoding and decoding schemes in Neural Networks dedicated to Speech processing tasks. As developed through section 2.1, several models have been proposed for mechanisms identified in the early stages of auditory processing. Our intention is to implement some of these models and propose variations to those, as a replacement to the simple convolutional encoder proposed in [1].

One first fundamental statement to make is that the role of the decoder is absolutely not symmetrical to that of the encoder in that approach: indeed, we want the network to produce an output directly comparable to its input to train it in a end-to-end supervised fashion. In that regard, the most intuitive scheme would be to design the decoder symmetrical to the encoder. But let us make a direct comparison to the human auditory system: our brain uses the encoded representation provided by the integrating modules of the early auditory system in order to perform Speech processing tasks, but it *does not perform synthesis of the target speech based on that representation*. This means that our equivalent human "network" is simply composed of the encoder and the separator network in fig.2.6.

However, as explained before, we need to design a decoder for supervised training purposes: making it symmetrical to the encoder would have no meaning in terms of biological inspiration, and indeed [1, 2] have shown that they obtained worse results when using the pseudo-inverse (symmetrical) of the learnt encoder as the decoder, than when using a jointly learnt convolutional decoder. From those results, and for the reasons explained here, we will focus on proposing an encoder structure, as we will present it in section 3, and simply use a learnt convolutional decoder.

Notice should also be made of the comparison between the so-called "Time-domain" algorithms as opposed to "Time-Frequency" ones. Several DNN-based Speech processing studies mention their algorithms to operate in the time-domain because they do not use an explicit Short-Time Fourier Transform - which in that case would be deemed a "Time-Frequency" representation. However they use encoder/decoder modules involving filterbanks, thus transforming the waveform to a subband-domain representation. We simply want to clarify that opposition by rather stating that:

- There are *fixed* representations like Cochleagrams, Mel-scale Spectrograms (or directly equivalently, Fourier spectrograms, hence the remark); and *learnt* representations like the outputs of a Convolutional layer for instance. Both these representations use filterbanks - explicitly in the first case and more implicitly in the second case - to obtain a representation of the waveform, which can be usually considered quite close to a Time-Frequency representation (because learnt filters often also display a bandpass structure).
- Additionally and maybe more importantly even, there are *magnitude*-based algorithms that operate only on the magnitude of the STFT representation (explaining why these are often lumped together with other subband-domain representations). This is to be opposed to algorithms which use versions of the signal without discarding an important part of the signal information which is the equivalent of the STFT *phase*.

This second point is of real importance, as it has been shown in [64] that Speech processing algorithms using the phase information as well as the magnitude information showed better results than those focusing only on magnitude. Phase was first ignored in most Speech Processing algorithms because it seemed too random to be easily subject to modelling.

Some solutions were provided on the one hand by focusing on phase-aware training targets like the Phase-Sensitive Mask or the complex-Ideal Ratio Mask [50], and showed promising results.

On the other hand, some studies focus on phase-aware training by directly adapting the structure of the encoder/decoder, which led to this branding of "Time-Domain" algorithms and was often related to Temporal Convolutional Networks [65, 66, 67, 61]. Closely related to our case, [68] shows in their experiments on Speech separation with Conv-TasNet that using the analytic version of the filters in the encoder (thus explicitly enforcing the use of the phase information) showed improvements in performances as compared to an encoder where only the real-part of the filters impulse responses were used - the comparison being made with an equivalent number of parameters.

Remark 2.1. *A filter is said to be analytic if its impulse response $u(t)$ respects the following condition 2.13:*

$$\mathcal{I}(u(t)) = \mathcal{H}[\mathcal{R}](u(t)) \tag{2.13}$$

that is if, its impulse response is a complex analytic function, which corresponds to the filter being shift-invariant to a delay in the time-domain. In [68], the "analytic" filterbank is obtained by concatenating the real-part (original) filters with their Hilbert transforms (the imaginary-part of the analytic extension).

Similarly, [2] replaces the convolutional encoder of Conv-Tasnet with a Multiphase Gammatone Filterbank, where each filter impulse response can be extended by a phase-shifted version of itself, with a phase factor ϕ. They find that their implementation outperforms Conv-TasNet for Speech separation on typical dataset WSJ0-2mix [69], when taking a number of phases equal to $N_\phi = 4$, that is, for $\phi \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$.

In that context, we can typically view the approach [68] applied to the Gammatone filterbank as being a case of a multiphase Gammatone filterbank, where $N_\phi = 2$ and $\phi \in \{0, -\frac{\pi}{2}\}$ (see 4.2).

3 Modelling Phase-Locking in an encoder for Speech Separation and Speech Enhancement with a TCN

3.1 Implementation Framework

For the following experiments, we use a TCN similar to that of Conv-TasNet [1] (see 2.2.3), where we simply replace the Encoder module, a learnt Convolutional layer with N filters, by a Gammatone-based custom encoder. As pointed before in subsection 2.2.4, we keep the learnt convolutional Decoder. The baseline structure is displayed on fig.3.1

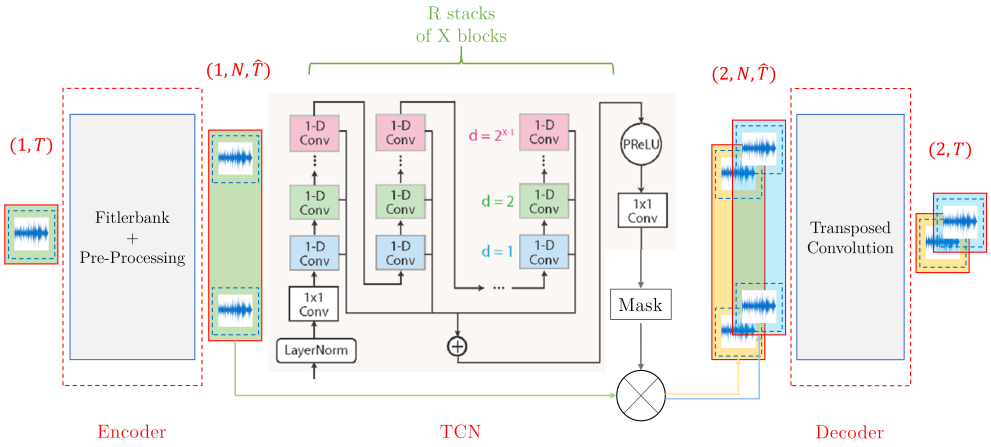


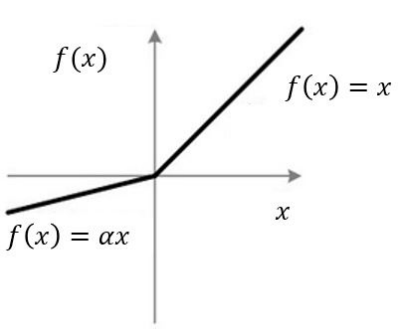
Figure 3.1: General TCN-based architecture used for Speech Separation.

3.2 Experiment objective

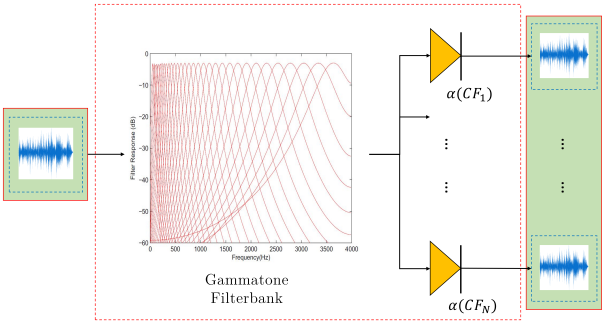
In this experiment, we will introduce a simple model to enable a Temporal Convolutional Network to learn Phase-Locking in each subband of a Gammatone filterbank. The encoder model takes into account the place-coding mechanism in the basilar membrane, and the possible time-coding mechanism by the early auditory nerve fibers.

The encoder model is described on fig.3.2b: the waveform enters the Gammatone Filterbank, where it is filtered by N frequency-tuned filters with the impulse response in eq.2.2. In each of the subband, the filtered waveform enters a Parametric ReLU, whose function graph is given on fig.3.2a: it is a unit with one trainable parameter representing the negative slope $\alpha \in \mathbb{R}$. Special parameter values are $\alpha = 1$, which transforms the PReLU into a simple identity function, $\alpha = 0$ which makes it equivalent to a classical ReLU performing half-wave rectification, and $\alpha = -1$ which takes the absolute value of the input (full-wave rectification).

With those PReLUs we are expecting to model frequency-dependent Phase-Locking: a parameter of $\alpha = 0$ would model a ReLU, that is, a phase-locked response, whereas $\alpha = 1$ would



(a) Function graph of PReLU: α is learnt over training.



(b) Encoder used in AudiTCN for the experiment.

mean no phase-locking is performed. In that view, we monitored the behaviour of $1 - \alpha(f)$, considering it would be a quantity following the direction of the variations of Vector Strength (closer to 1 for phase-locked responses and closer to 0 for unmodified response), we will refer to this quantity as *pseudo-Vector Strength* (pVS) in the following.

The objective of the experiment is then to answer the following questions: does the Neural Network optimize its parameters to learn a time-coding mechanism via Phase-Locking? And if so, is the frequency-dependency of Phase-Locking related to what is observed in the human and animal? Or is this organization exclusive to the human auditory system and determined by other factors than sole optimization perspective?

Remark 3.1. *To answer those questions, we did not take the optimized hyper parameters in [1] because this would result in a model with more than 8M parameters, for an optimal performance we do not need. We instead used lighter versions with between 1.2M and 3.9M parameters, for faster training.*

We train first the proposed network (we will call it *Auditory Temporal Convolutional Network* (AudiTCN) for convenience) toward a Speech Separation objective on the WSJ0-mix2 dataset [69] comprising a training set (30h), a cross-validation set (10h) and a test set (5h). The data is composed of utterances from 2 speakers (we take the speaker-independent version, meaning the speakers are not all the same throughout the different sets) reading the Wall Street Journal, and mixed at a random SNR between 0 and 10 dB, at a sampling frequency of 8kHz in our case.

We then make the comparison (see subsection 3.3) with the same network in a Speech Enhancement setting: we simply consider one of the speaker to be the noise source. The clean speech dataset used consists in 6'000 utterances from WSJ-0 resampled at 8kHz with no windowing. The noise dataset is composed of 15 non-stationary noises from the NOISEX corpus: 11 for the training set, 2 for the cross-validation set and 2 for the testing set. The noise samples being more than 4 minutes long, for each clean sample, we pick a random starting point in a randomly chosen noise sample, extract the noise chunk and mix it to the clean sample with a SNR randomly chosen between $-5dB$ and $5dB$. 10 % of the training mixtures are only composed of noise, to emphasize on noise modelling [70, 63]).

In both separation and enhancement cases, the objective cost function is Scale-Invariant SNR [71] a simpler alternative to the classical SDR (Signal-to-Distortion ratio) of the BSS evaluation

toolkit [72], which takes into the account the fact that target speech and interference might not be correctly relatively scaled. In separation, Permutant-Independent Training [73] is applied to address the permutation problem of blind speech separation.

We simply take classical SNR and SI-SNR as evaluation metrics, as our goal is not directly to show possible improvements on perceptual results with this kind of experiment: therefore, the evaluation is mainly dedicated to verifying the network is not over-fitting to the training data.

3.3 Results

In the first sub-experiment, we compare on a Speech Separation task our encoder with free PReLU (*AudiTCN*) with a similar network with the PReLU fixed in a ReLU setting (*AudiTCN-ReLU*) and in a identity setting (*AUdiTCN-ID*). Hyper parameters are found in table 3.1, and performance results in table 3.2. A learning rate of 10^{-4}, a batch size of 3 and 50 epochs were used.

N	Number of frequency subbands	256
R	Number of dilated convolutional stacks	2
X	Number of convolutional blocks per stack	6
B	Filters in pointwise convolutions (after Bottleneck)	256
H	Filters in depthwise convolutions	512
P	Kernel size in depthwise convolutions	3

Table 3.1: AudiTCN hyper parameters for first sub-experiment

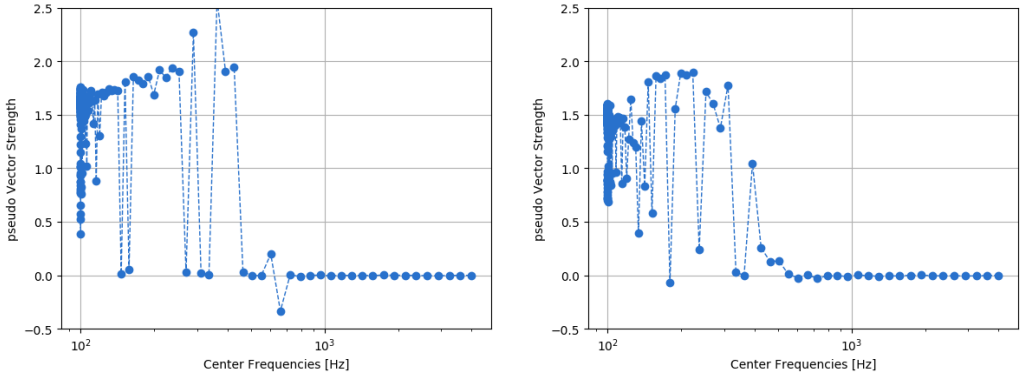
	AudiTCN-ID	AudiTCN-ReLU	AudiTCN
Number of parameters	3'391'000	3'391'000	3'391'256
SDR improvement (dB)	12.82	11.66	**13.14**
SI-SDR improvement (dB)	12.49	11.28	**12.82**

Table 3.2: Performances comparison between AudiTCN variants. Best performance is indicated in bold.

We visualize PReLU behaviour via pseudo-VS as a function of the subband center-frequency in fig.3.3a: we observe that the distribution of pVS tends to a binary distribution where $pVS \in \{0, 2\}$, alternating between the two values for frequencies up to $500Hz$, and $pVS = 0$ consistently for frequencies higher than $500Hz$. The similar behaviour was observed in independent experiments.

We repeated the same experiment on a Speech Enhancement task, with only the AudiTCN network: the observed behaviour is exactly similar (see 3.3b), indicating the optimization of that PReLU module is task-independent.

The second part of the experiment aims at studying the role of the filterbank in the behaviour of the PReLU module: we use a modified 128-channel filterbank where the Center-Frequencies CF_k are uniformly distributed on the Hertz scale, and a constant bandwidth parameter b_k is fixed. We use the Gammatone impulse response (eq.2.2). We compare two of these filterbanks: one has a filters with large bandwidth, the other with small bandwidth. The task is Speech Separation. The resulting pseudo-Vector Strength values are plotted on fig.3.5 next to their



(a) pVS as a function of Center-Frequencies in AudiTCN trained on the Speech Separation task.

(b) pVS as a function of Center-Frequencies in AudiTCN trained on the Speech Enhancement task.

Figure 3.3: Results of first sub-experiment: quasi-binary distribution of pseudo-Vector Strength.

corresponding filterbank. We observe that the randomly alternating pseudo-Vector Strength behaviour is occurring on frequency ranges where filters largely overlap, whereas for more separated filters, pVS is a steady 0.

Finally in a third, complementary sub-experiment, we investigate the role of the initial PReLU parameter α_0 in the observed behaviour. We take a 256-channel uniform filterbank with Gammatone impulse responses as in the previous sub-experiment, with a large bandwidth, and compare results for $\alpha_0 \in \{-1, -0.3, 0, 0.3, 1\}$. Resulting pseudo-VS values are plotted on 3.6. We observe that the initialization of α plays a role in the final pVS values: a distribution initialized at $\alpha_0 = 1$ will not be modified during training, whereas intermediary initializations ($\alpha_0 \in \{-0.3, 0.3\}$) will produce the most diverse pVS final values. Initialization of $\alpha_0 = -1$ produces also an alternating distribution, but the training is somehow slower, making the distribution take more intermediary values than expected of a final state (convergence is not attained at the 50^{th} epoch, as opposed to the other experiments).

3.4 Discussion

First sub-experiment

The results of the first sub-experiment are indicative of several aspects:

First, the values of the PReLU parameter α (and equivalently, the pseudo-Vector Strength) are almost binary: $\alpha \in \{-1, 1\}$. This seems to indicate that these two settings - corresponding to full-wave rectification and identity, respectively - are optimal in representing the information in this encoding framework. The value $\alpha = -1$ is particularly disturbing in our study, as we can not relate it to biological meaning. Indeed, the firing activity being a non-negative encoding, a full-wave rectified input is totally equivalent to the same unmodified input, when viewed by a the first auditory neurons making the transduction from soundwave to impulses. In that respect, it seems the network does not train in the direction of bringing non-linear transformation to the subband signals.

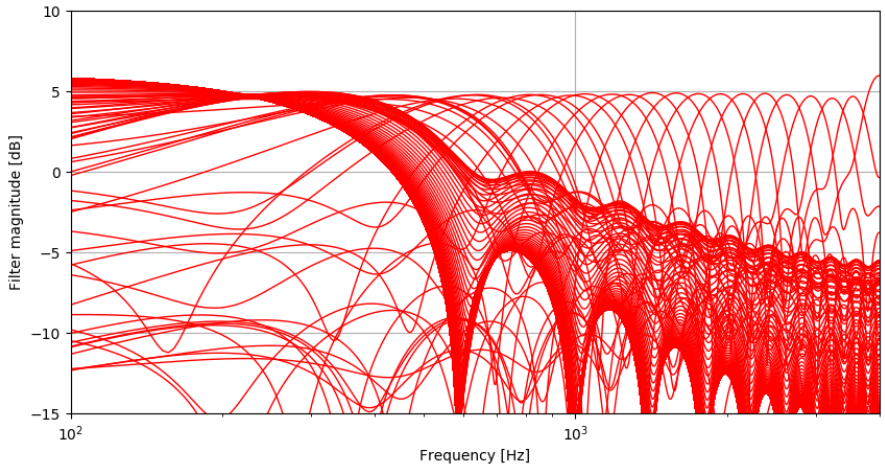


Figure 3.4: 256-channel Gammatone filterbank used in first both Speech Separation and Speech Enhancement experiments.

Secondly, pVS randomly alternates between its two values for $CF < 500Hz$, and becomes steady at 0 for $CF > 500Hz$. This behaviour seems also not to be task specific, although we explained earlier that Speech Separation and Speech Enhancement do not focus on the same frequency ranges. This seems to indicate that the observed frequency-dependency of pVS is *a priori* rather determined by the structure of the encoder itself, than the subsequent separator network.

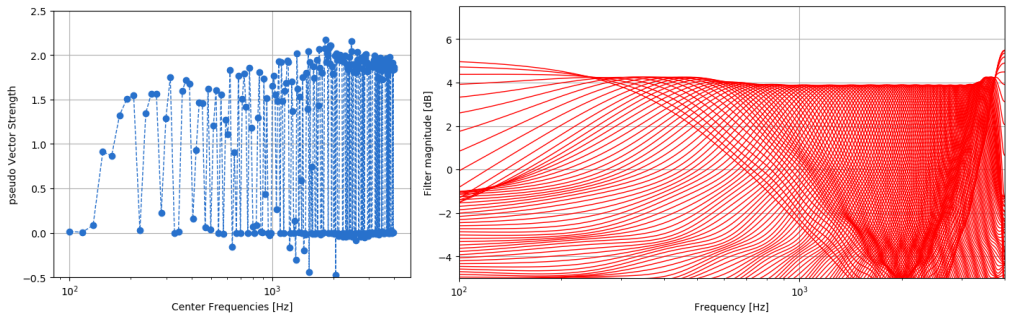
In that latter view, we visualized the Gammatone filterbank in both the Speech Separation and Speech Enhancement experiments (see fig.2.1c). We noticed that since the number of filters used is quite large ($N = 256$), the limited filter length ($L = 20$) does not suffice to produce filters with acute frequency-tuning in the low-frequencies. Therefore, given the mentioned phenomenon and the non-linear ERB-scale distribution of Center-Frequencies, filters largely overlap in the low-frequencies - for $CF < 500Hz$ - which coincides with the observed randomly alternating pVS observation.

From those observations, we make the following assertions:

1. The Neural Network does not learn anything equivalent to Phase-Locking, not with regard to the values of pseudo-Vector Strength nor with regard to the observed frequency-dependency.

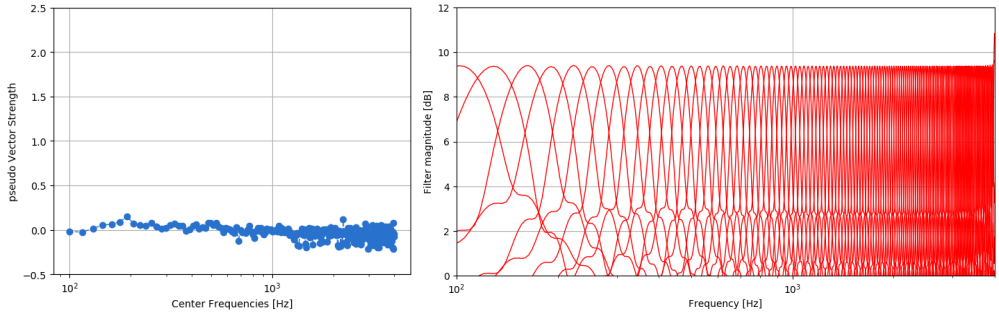
2. The network tries to produce different versions of the same information where that information is redundant (in the case of that experiment in the low-frequency range where filters largely overlap, thus outputting similar information from one filter to the neighbouring one). Those different versions are very close (one is the original subband waveform, the other a full-wave rectified version of that waveform) and do not suppress information, therefore we consider those as *distortionless*.

In addition, when putting a ReLU at the output of the encoder, this distortionless propriety is not respected anymore, and we see that the performance is significantly suboptimal yielding a loss of nearly 1.5 dB SI-SNR (and SNR) improvement (see tab.3.2) when compared to the identity and PReLU cases. With regard to the previous observations, we understand it to be the result



(a) pVS as a function of Center-Frequencies for AudiTCN with large-bandwidth filterbank.

(b) 128-channel uniform filterbank with large bandwidth used in AudiTCN.



(c) pVS as a function of Center-Frequencies for AudiTCN with small-bandwidth filterbank.

(d) 128-channel uniform filterbank with small bandwidth used in AudiTCN.

Figure 3.5: Results of second sub-experiment: pseudo-VS distribution changes according to the relative separation of filters.

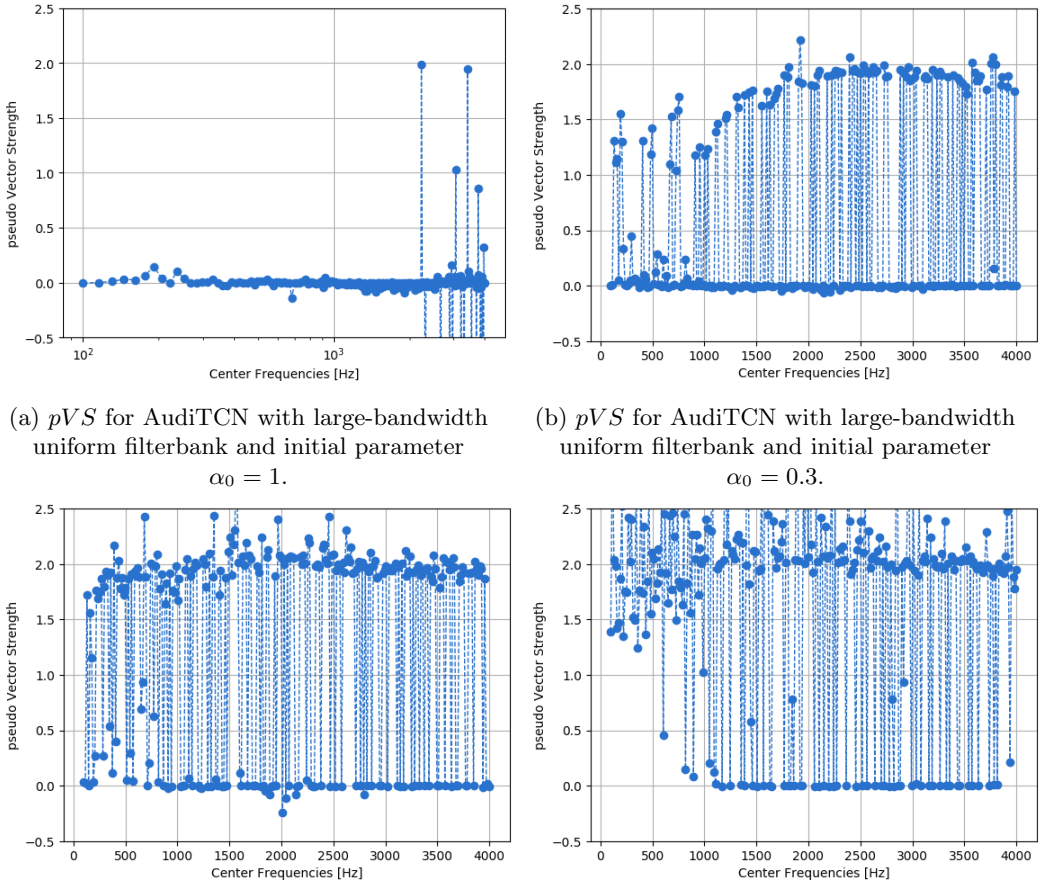
of losing one half of the information, which can be problematic if the frequency-tuning of filters is not very acute allowing for fast variations of the signal to happen during the negative part of the waveform. The AudiTCN-PReLU version performed almost similarly to AudiTCN-ID, with a difference of only 0.3 dB for both metrics, which is not significant given the fluctuations of the validation loss after convergence.

Remark 3.2. *This does not go against the results obtained in [1], where authors study the influence of the ReLU at the output of the Conv-TasNet encoder and conclude they do not observe any difference caused by the presence or absence of ReLU. The encoder being a learnt convolutional filterbank, it can adapt to the presence of this ReLU to still yield an optimal configuration, by consequently modifying the filters structures.*

Second sub-experiment

Results from this second sub-experiment help us verifying the previous assertion regarding the adaptative behaviour of the Neural Network with respect to the quantity of redundant information available.

We observe that when filters largely overlap (see fig.3.5b) the pVS values randomly alternate within the binary distribution (see fig.3.5a), this without any frequency dependency since we



(a) pVS for AudiTCN with large-bandwidth uniform filterbank and initial parameter $\alpha_0 = 1$.

(b) pVS for AudiTCN with large-bandwidth uniform filterbank and initial parameter $\alpha_0 = 0.3$.

(c) pVS for AudiTCN with large-bandwidth uniform filterbank and initial parameter $\alpha_0 = -0.3$.

(d) pVS for AudiTCN with large-bandwidth uniform filterbank and initial parameter $\alpha_0 = -1$.

Figure 3.6: Results of third sub-experiment: pseudo-VS distribution changes according to the initialization of the PReLU parameter α_0.

used a filterbank with constant bandwidth across uniformly distributed center-frequencies. On the other hand, when filters are more separated (see fig.3.5d), the pVS values converge to a steady 0 distribution (see fig.3.5c), meaning the optimal behaviour in that case, is simply an identity mapping of each subband waveform.

This confirms the previous assertion that it in this fixed filterbank encoder, subband information is diversified if redundancy is present, and kept untouched otherwise.

Third sub-experiment

The results in fig.3.6 show the initialization of the PReLU parameter α plays an important role in the training of the network: so far we initialized our PReLU units in a ReLU setting ($\alpha_0 = 0$). Since we observed earlier than both $\alpha = -1$ and $\alpha = 1$ seem to be local optima for a single PReLU unit, if we initialize the PReLU parameters too close to one of these parameters,

the values will stay in this local optimum and will not output the optimal distribution in the large-bandwidth case, that is, an alternating distribution. This is especially true for $\alpha_0 = 1$, but less so for $\alpha_0 = -1$: the PReLU values in that latter case still converge to an alternating distribution, but at a slower pace, indicating that $\alpha = -1$ seems to be a less optimal value than $\alpha = 1$, forcing the network to train out of this setting.

The intermediary distributions ($\alpha_0 \in \{-0.3, 0, 0.3\}$) draw the PReLU parameters far enough from the local attractors ± 1, enabling the network to learn a global distribution (aternating in the large-bandwidth scenario) which is optimal compared to what is obtained with the previously cited initialization.

4 Multiphase extension of a Gammatone-based encoder for Speech Separation with a TCN

4.1 Implementation Framework

The baseline implementation is exactly the same as in the previous section, that is, an encoder based on filterbank decomposition, a TCN separator network and a learnt convolutional decoder as in TasNet: see fig.3.2b.

4.2 Experiment objective

As indicated in subsection 2.2.4, phase information can be explicitly exploited in an encoder by either considering an analytic extension of the filterbank [68], or introducing phase-shifts in the filter impulse response to produce different phase-shifted versions of the same filter [2]. We relate these approaches to the results presented in the previous section, where it was shown that a Neural Network will make the best out of information redundancy by producing different representations of that same information.

Remark 4.1. *We usually refer to both works [68, 2] together as in our Gammatone filterbank case, the analytic extension in [68] reduces to a multiphase case of [2] where the phase factors are* $\phi \in \{0, -\frac{\pi}{2}\}$ *since:*

$$h_k^{\mathbb{C}}(t) = at^{n-1}e^{j2\pi CF_k^{(ERB)}t}e^{-2\pi b_k t} = h_k^{\mathcal{R}}(t) + jh_k^{\mathcal{I}}(t) \tag{4.1}$$

$$\begin{cases} h_k^{\mathcal{R}}(t) = at^{n-1}\cos(2\pi CF_k^{(ERB)}t)e^{-2\pi b_k t} \\ h_k^{\mathcal{I}}(t) = at^{n-1}\cos(2\pi CF_k^{(ERB)}t - \frac{\pi}{2})e^{-2\pi b_k t} = at^{n-1}\sin(2\pi CF_k^{(ERB)}t)e^{-2\pi b_k t} \end{cases} \tag{4.2}$$

Here, we present the Neural Network with the possibility to model different representations by passing the signal through phase-shifted versions of each filter in the fixed filterbank, where the phase factors are trainable parameters (see schematics on fig.4.1). The operation performed by this encoder is then the following, with $\otimes$ denoting the Kroenecker product, $K = \lfloor \frac{N}{N_\phi} \rfloor$ being the number of frequency channels, N_ϕ being the number of phase shifts, $x \in \mathbb{R}$ the input and $\mathbf{z} \in \mathbb{R}^N$ the output:

$$\mathbf{z} = x * \begin{bmatrix} h_1^{\mathcal{R}} & h_1^{\mathcal{I}} \\ \vdots & \vdots \\ h_K^{\mathcal{R}} & h_K^{\mathcal{I}} \end{bmatrix} \otimes \begin{bmatrix} \cos\phi_1 & \dots & \cos\phi_{N_\phi} \\ \sin\phi_1 & \dots & \sin\phi_{N_\phi} \end{bmatrix} \tag{4.3}$$

Remark 4.2. *The operation would be more straight-forward in complex representation, using the complex version of the Gammatone impulse response* $h_k^{\mathbb{C}}$ *in eq.4.3, but PyTorch framework does not allow yet for complex numbers to be used in tensors supported on CUDA architecture. For that reason, we use the real- and imaginary-part representation for the operation.*

We train our resulting network (called *Multiphase Temporal Convolutional Network* (PhiTCN) for convenience) on a Speech Separation objective, with the hyperparameters found in table 4.1. The network used is lighter in its number of parameters, our objective being initially to monitor the evolution of the learnt phase-shift factors over the training. We initialize the phase-shifts

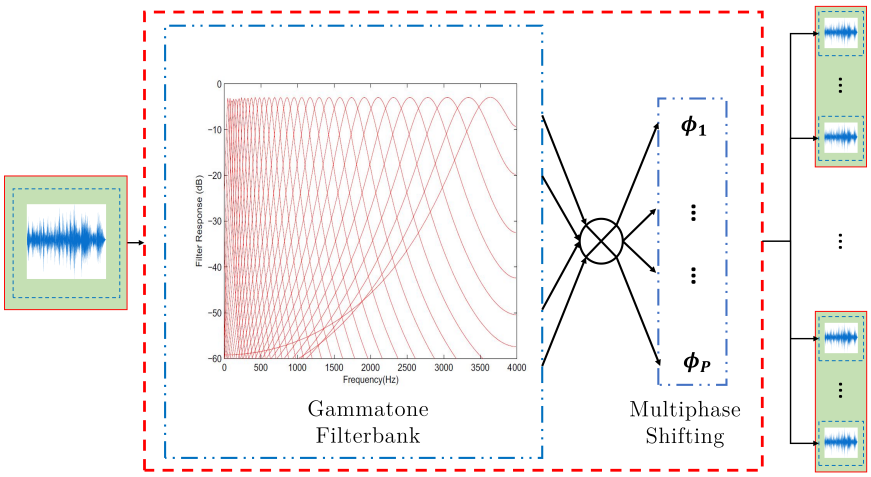


Figure 4.1: Schematics of the proposed parametric Multiphase Gammatone-based encoder.

(random uniform distribution on $[0, 2\pi]$) and monitor the evolution of phase factors for 6 independent experiments.

N	Subbands at the encoder output	256
N_ϕ	Phase factors	4
K	Frequency subbands	64
R	Dilated convolutional stacks	1
X	Convolutional blocks per stack	5
B	Filter in pointwise convolutions	256
H	Filters in depthwise convolutions	512
P	Kernel size in depthwise convolutions	3

Table 4.1: PhiTCN hyper parameters

Finally, we combine both approaches by adding PReLU at the output of each channel in the PhiTCN encoder (we call it *Phi-P-TCN*), and train a heavier version of the network, with the hyperparameters found in table 4.2 (same as in [1]). We compare Speech Separation performances with the baseline Conv-TasNet provided in [1], a simple Gammatone-based TCN and our approach.

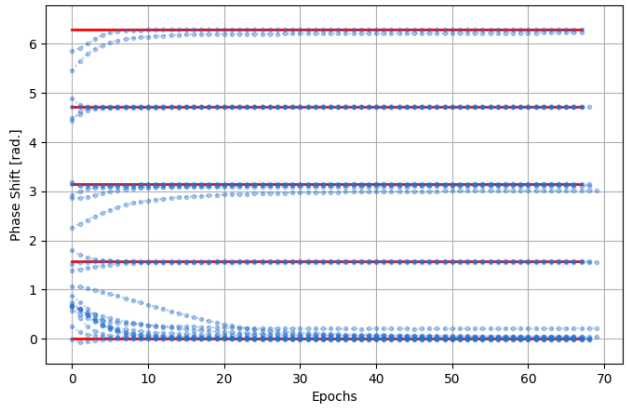
4.3 Results

Phase distribution

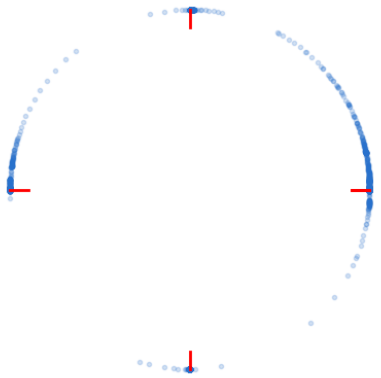
The evolution of phase factors over the training is plotted on fig.4.2a. We can clearly identify that the phase factors distribution converge to a discrete distribution where the factors are of the form $\phi_p = \frac{2\pi p}{N_\phi}$, which leads to a uniform distribution on the unit circle for the complex representation corresponding to those phase factors (see fig.4.2b).

		Conv-TasNet	Gammatone-TCN	Phi-P-TCN
N	Subbands at the encoder output	512	512	512
N_ϕ	Phase factors	N/A	N/A	8
K	Frequency subbands	512	512	64
R	Dilated convolutional stacks	4	4	4
X	Convolutional blocks per stack	8	8	8
B	Filters in pointwise convolutions	256	256	256
H	Filters in depthwise convolutions	512	512	512
P	Kernel size in depthwise convolutions	3	3	3

Table 4.2: Network Hyperparameters for the comparison between Conv-TasNet, Gammatone-TCN and Phi-P-TCN.



(a) Phase-shift factors ϕ_p learnt over 6 independent trainings of the proposed PhiTCN model. Red bars indicate the discrete uniform distribution.



(b) Complex representation of the learnt $e^{j\phi_p}$ phase factors on the unit circle. Red ticks indicate the discrete uniform distribution

Figure 4.2: Results of second experiment, showing convergence of the phase-shift factors over training for independent sub-experiments.

Final comparison

Results of the comparison between classical Conv-TasNet, Gammatone-based TCN and our approach Phi-P-TCN are displayed in table 4.3. Best performance is obtained for PHi-P-TCN.

	Conv-TasNet	Gammatone-TCN	Phi-P-TCN
Number of parameters	8'918'080	8'907'840	8'908'360
SDR improvement (dB)	15.7	15.7	**16.4**
SI-SDR improvement (dB)	15.4	15.4	**16.1**

Table 4.3: Results for comparison between Conv-TasNet, Multiphase-Gammatone TasNet and Phi-P-TCN. Best performance is indicated in bold

4.4 Discussion

We interpret the previous results as typical strategy of a neural network optimizing for maximally discriminative analysis power. Indeed, the discriminative measure of the analysis is directly linked to the cumulated distance between parameters of the analysis (in our case, the phase-factors), and this distance is maximized in a uniform distribution.

This justifies the intuition in [2], and we might add that compared to their approach, we did not enforce the symmetry with respect to π as we were not using a ReLU at the output of the encoder, but the resulting distribution is still symmetrical with respect to π nonetheless.

We did not consider odd numbers of phase factors: as stated in the previously cited work, it might be suboptimal if we put a ReLU at the output of the encoder (because we would lose the negative part of the waveform, without recovering it with the corresponding negative filter), but without ReLU in the encoder as in our work, we believe it would not make any difference since we do not need the symmetry with respect to π for the phase distribution to be optimal.

In conclusion, we can replace the learnt encoder of Conv-TasNet with a fixed multiphase-Gammatone filterbank, thus reducing the number of parameters without sacrificing optimality as we know that:

- The convolutional weights converge to a Gammatone filterbank when training on speech data (from [1, 2])
- From our contribution, the multiphase approach yields some additional discriminative power, and it is optimized when the phase factors are uniformly distributed on the $[0, 2\pi]$ range.

Finally we notice that with the Phi-P-TCN approach we gain 0.7dB in both SDR and SI-SDR improvements compared to a classical COnv-TasNet or a Gammatone-based TCN. The performance is very similar to the same configuration without trainable phase shifts found in [2], which is understandable since we noticed earlier that the phase shifts tend to a uniform distribution on the unit circle, which is the fixed configuration used in that work.

5 Future work: Modelling correlation variations in cortical layers

5.1 Correlated variability in cortical circuits

In the previous experiments, we observed some interesting behaviour when studying the overlapping between different filters, which we assumed was able to model frequency-tuned units (hair cells) with a certain degree of correlation. This correlation is thoroughly studied and modelled in [12] where a model is proposed to fit human experimental data on frequency and intensity discrimination. The model is composed of virtual neural units tuned to a Center-Frequency, the firing activity of the organ being modelled by a multivariate Gaussian random process. As opposed to [28] which considered a model with independent processing across channels, here the correlation across channels is modelled by considering a non-diagonal covariance matrix Σ. We give the following expression for the degree of correlation between two frequency-tuned units' i and j firing activities:

$$\Sigma_{i,j} = \rho_{i,j}\sqrt{\mathbb{E}_i\mathbb{E}_j} \tag{5.1}$$

Remark 5.1. *The underlying model chosen for firing activity statistics is often a Poisson process, resulting in the variance of a unit's firing activity being taken equal to its mean, which is the case here in eq.5.1.*

This modelling of correlation stems from studies on the visual primary cortex [74, 75]: the correlation parameters $\rho_{i,j}$ are taken between 0 and a maximal value ρ chosen empirically. It is also referenced in [75] that this parameter is lower in input cortical layers (they refer to a higher discrimination power, which is synonym with lower correlation across units), indicating there are various spectral scales of the correlation between neural units in the auditory periphery system, introducing a *correlated variability*. They perform an analysis of a macaque visual primary cortex firing activity, tracking the correlation between spike counts in two different groups of layers: a granular layer (closer to the input cortical layers) and infra-granular or supra-granular layers (closer to the output cortical layers). Theirs results indicate that:

- The correlation between spike counts is overall higher in the infra- and supra-granular layers, which they account for by noticing that the spatial spread of cortical connections in these layers is larger than in the granular layer, enabling for more spread information to be integrated. By contrast, the information processed in the granular layer is mostly local, therefore the response is uncorrelated for distant neurons.
- The spike-count correlation does not depend on tuning similarity in the granular layer (the tuning similarity in visual cells is the preferred orientation, whereas in our auditory case, this would rather be the center-frequency), whereas it increases with tuning similarity for supra- and infra-granular layers.

In that view, we can conclude that information needs to be largely shared for output layers because their role is to integrate different projections from earlier layers, whereas input layers need to have a high discriminative power, which can only come from a local processing of the information, therefore reducing the degree of correlation between neighbouring units.

5.2 Modelling correlated variability in Temporal Convolutional Networks

From these observations, it would be interesting to consider a TCN implementation where this correlation between neighbouring units can be taken into account and quantified. We propose to use *Grouped Convolutional layers* to encode this correlation: Grouped Convolutions are convolutions which regroup some of the computations (see fig.5.1). This means that instead of having one matrix with $C_{in} \times C_{out}$ filters in it as in a regular convolution, there are C_{groups} matrixes each having $\frac{C_{in}}{C_{groups}} \times \frac{C_{out}}{C_{groups}}$. The operations corresponding to Regular and Grouped Convolutions are:

$$\begin{cases} \text{Regular Convolution} & y_k = \sum_{j=0}^{C_{in}-1} x_j * h_{j,k} \\ \text{Grouped Convolution} & y_k = \sum_{j=0}^{\lfloor \frac{C_{in}}{C_{groups}} \rfloor - 1} x_{j+l\times\lfloor \frac{C_{in}}{C_{groups}} \rfloor} * h^{(l)}_{j,k-l\times\lfloor \frac{C_{out}}{C_{groups}} \rfloor} \end{cases} \tag{5.2}$$

with the *group index* $l = \lfloor \frac{k}{\lfloor \frac{C_{out}}{C_{groups}} \rfloor} \rfloor$

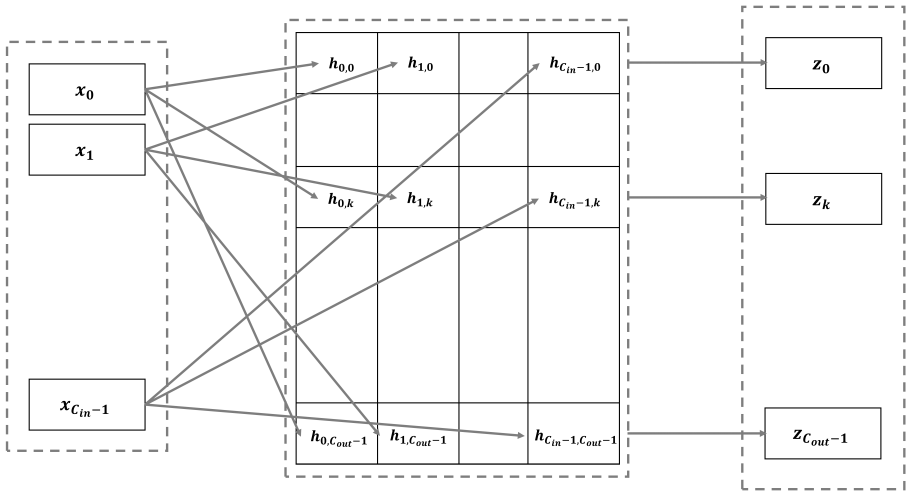
We assume the relation of those operations to the spectral correlation mentioned before to be the following:

- Uncorrelated units ($\rho = 0$) have a processing which is equivalent to what is called a Depthwise Convolutional layer: each input channel has its own group and corresponding filter (extreme example of Grouped Convolution in which $C_{groups} = C_{in}$).
- Highly correlated units (ρ close to 1) are closer to regular Convolutional layer: all the input channels share the same bank of filters ($C_{groups} = 1$), which is similar to considering long-range horizontal circuitry as mentioned in [75].

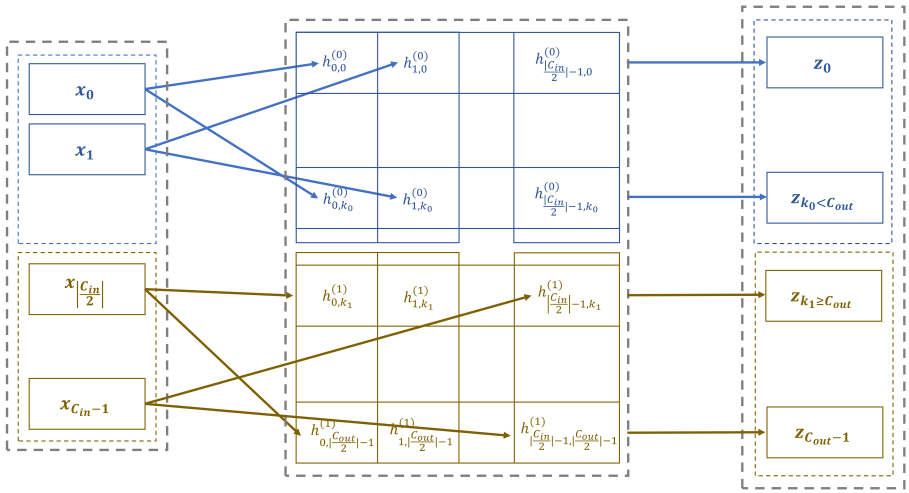
We sketch a variant of TCN based on the proposed idea on fig.5.2, where g is the number of groups in each convolution of the "1D-Conv" blocks (which are modified into Grouped Convolutions). In that architecture, each stack of convolution blocks is considered as equivalent to a cortical layer: in that view, input layers have the highest number of groups, corresponding to weak correlation across channels, while output layers have the smallest number of groups, corresponding to stronger correlation.

In that sense, we directly quantify the degree of correlation between units by the number of groups g used in Grouped Convolutional layers. The corresponding objectives and points of interest include:

- Reducing the number of parameters in the Neural Network, making it easier to train and less prone to over-fitting. Indeed, the number of parameters in a Grouped Convolutional layer with $C_{groups} < C_{in}$ groups, filters of length L and C_{out} output channels is $L \times \frac{C_{in}}{C_{groups}} \times C_{out}$, versus $L \times C_in \times C_{out}$ for a classical Convolutional layer.
  This would enable us with optimization tools for network complexity based on physiological understanding.
- Investigating more thoroughly the scale variability of spectral correlation observed in the visual cortex in [74] for the auditory case.



(a) Regular convolution schematics.



(b) Grouped convolution schematics.

Figure 5.1: Comparison between Grouped and Regular Convolutions. The time dimension is not represented here

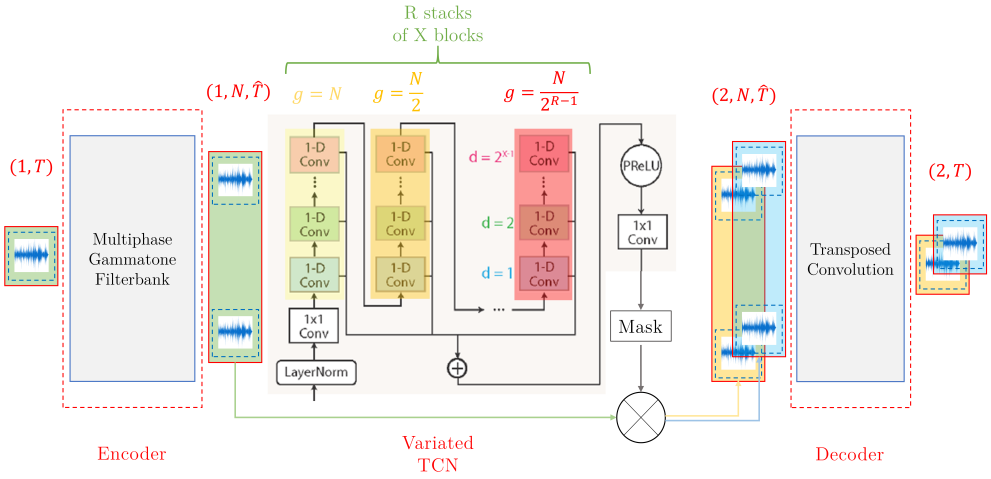


Figure 5.2: Proposed TCN Variant, taking into account a variation in the correlation between convolution channels

Conclusion

We introduced an encoder model for speech processing with Temporal Convolutional Networks, based on place- and time-coding schemes identified in the human auditory system. When training those networks on speech data for Source Separation and Enhancement, we observed that, although a learnt convolutional encoder had been proved to mimick the human place-coding mechanism by converging to a Gammatone filterbank configuration [1, 2], our model was not able to reproduce the phase-locking phenomenon used for time-coding in humans.

However, further study of the behaviour of the Parametric Rectifying units (PReLU) introduced in our model led us to understand that the Neural Network tended to produce diverse representations of the same information when redundancy was present. The said representations were found to be either an identity mapping of the subband waveform, or a full-wave rectified version of it, which is disortionless and identical from a biological point of view - the resulting firing activity being identical when feeding either of these inputs to an auditory neuron. We interpreted this behaviour as the encoder strategy to be discriminative in its analysis without distorting the information, which can be possibly linked to its early position in the audio processing network.

As a consequence and inspired from previous works [68, 2], we proposed a multiphase extension of our Gammatone-based encoder for Speech separation and enhancement with a TCN. The model had the ability to learn optimal phase-shift factors to exploit redundant phase information in a Gammatone filterbank. Experiments showed that the optimal distribution learnt for those phase factors was a discrete uniform distribution on the unit circle. This was again interpreted as a strategy maximizing the discriminative power of the encoder analysis, motivating the intuition in [2].

We then compared performances between the classical convolutional learnt encoder, a fixed Gammatone-based encoder and our parametric multiphase Gammatone-based encoder on a Speech Separation task. We noticed that best performance was achieved for the multiphase approach, but that the Parametric rectifying did not bring any further improvement.

Finally, we rapidly introduced ideas of a strategy for reducing the number of parameters in a TCN without sacrificing performance, by exploiting a layer-dependent correlation measure in the convolution operations of a TCN. This idea is based on single unit recordings in different layers of the primary visual cortex [74, 75], showing different levels of information integration with respect to the considered layer.

References

[1] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, 2019.

[2] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via tasnet," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 36–40, 2020.

[3] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," 01 1988.

[4] B. Gorman, *A Framework for Speechreading Acquisition Tools*. PhD thesis, 03 2018.

[5] M. Rudnicki, O. Schoppe, M. Isik, F. Völk, and W. Hemmert, "Modeling auditory coding: from sound to spikes," *Cell and Tissue Research*, vol. 361, pp. 159 – 175, 2015.

[6] D. Havelock, S. Kuwano, and M. Vorländer, *Handbook of signal processing in acoustics*. Springer Science & Business Media, 2008.

[7] C. Schreiner and G. Langner, "Periodicity coding in the inferior colliculus of the cat. ii. topographical organization.," *Journal of neurophysiology*, vol. 60 6, pp. 1823–40, 1988.

[8] R. Brette and D. F. M. Goodman, "Simulating spiking neural networks on gpu," *Network: Computation in Neural Systems*, vol. 23, pp. 167 – 182, 2012.

[9] D. Ditter and T. Gerkmann, "Influence of speaker-specific parameters on speech separation systems," in *INTERSPEECH*, 2019.

[10] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

[11] T. Zhang and J. Wu, "Discriminative frequency filter banks learning with neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, pp. 1–16, 2019.

[12] C. Micheyl, P. R. Schrater, and A. J. Oxenham, "Auditory frequency and intensity discrimination explained using a cortical population rate code," *PLoS Computational Biology*, vol. 9, 2013.

[13] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *INTERSPEECH*, 2010.

[14] P. Johannesma, *The pre-response stimulus ensemble of neurons in the cochlear nucleus*. Hearing : international symposium, Instituut voor Perceptie Onderzoek (IPO), 972.

[15] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[16] J. E. Rose, J. F. Brugge, D. Anderson, and J. Hind, "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey.," *Journal of neurophysiology*, vol. 30 4, pp. 769–93, 1967.

[17] A. Møller, "Frequency selectivity of phase-locking of complex sounds in the auditory nerve of the rat," *Hearing Research*, vol. 11, pp. 267–284, 1983.

[18] E. Verschooten, S. Shamma, A. Oxenham, B. Moore, P. Joris, M. Heinz, and C. Plack, "The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints," *Hearing Research*, vol. 377, pp. 109 – 121, 2019.

[19] J. L. van Hemmen, "Vector strength after goldberg, brown, and von mises: biological and mathematical perspectives," *Biological Cybernetics*, vol. 107, pp. 385–396, 2013.

[20] T. Weiss and C. Rose, "A comparison of synchronization filters in different auditory receptor organs," *Hearing Research*, vol. 33, pp. 175–179, 1988.

[21] S. Kale and M. G. Heinz, "Envelope coding in auditory nerve fibers following noise-induced hearing loss," *Journal of the Association for Research in Otolaryngology*, vol. 11, pp. 657–673, 2010.

[22] D. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones.," *The Journal of the Acoustical Society of America*, vol. 68 4, pp. 1115–22, 1980.

[23] C. P. C. Versteegh, S. W. F. Meenderink, and M. van der Heijden, "Response characteristics in the apex of the gerbil cochlea studied through auditory nerve recordings," *JARO: Journal of the Association for Research in Otolaryngology*, vol. 12, pp. 301 – 316, 2010.

[24] A. N. Temchin and M. A. Ruggero, "Phase-locked responses to tones of chinchilla auditory nerve fibers: Implications for apical cochlear mechanics," *Journal of the Association for Research in Otolaryngology*, vol. 11, pp. 297–318, 2009.

[25] A. Brughera, L. Dunai, and W. M. Hartmann, "Human interaural time difference thresholds for sine tones: the high-frequency limit.," *The Journal of the Acoustical Society of America*, vol. 133 5, pp. 2839–55, 2013.

[26] B. C. J. Moore and S. M. A. Ernst, "Frequency difference limens at high frequencies: evidence for a transition from a temporal to a place code.," *The Journal of the Acoustical Society of America*, vol. 132 3, pp. 1542–7, 2012.

[27] A. J. Oxenham, C. Micheyl, and M. V. Keebler, "Can temporal fine structure represent the fundamental frequency of unresolved harmonics?," *The Journal of the Acoustical Society of America*, vol. 125 4, pp. 2189–99, 2009.

[28] T. Dau, "Modeling auditory processing of amplitude modulation," *Journal of the Acoustical Society of America*, vol. 101, pp. 3061–3061, 1997.

[29] J.-H. Lestang and D. F. M. Goodman, "Canonical brain computations account for perceived sound source location," *bioRxiv*, 2019.

[30] E. Lopez-Poveda and A. Eustaquio-Martín, "A biophysical model of the inner hair cell: The contribution of potassium currents to peripheral auditory compression," *Journal of the Association for Research in Otolaryngology*, vol. 7, pp. 218–235, 2006.

[31] A. Palmer and I. Russell, "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells," *Hearing Research*, vol. 24, pp. 1–15, 1986.

[32] S. O. Rizzoli and W. J. Betz, "Synaptic vesicle pools," *Nature Reviews Neuroscience*, vol. 6, no. 1, pp. 57–69, 2005.

[33] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[34] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[35] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[36] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, pp. 2403–2418, 2001.

[37] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source–filter-based single-channel speech separation using pitch information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 242–255, 2011.

[38] M. Wohlmayr, M. Stark, and F. Pernkopf, "A mixture maximization approach to multipitch tracking with factorial hidden markov models," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5070–5073, 2010.

[39] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 799–810, 2011.

[40] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2299–2310, 2007.

[41] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, pp. 177–180, 2003.

[42] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *ICA*, 2004.

[43] M. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *INTERSPEECH*, 2006.

[44] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3749–3753, 2014.

[45] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4029–4032, 2008.

[46] Y. Wang and M. Brookes, "Model-based speech enhancement in the modulation domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 580–594, 2018.

[47] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 466–475, 2003.

[48] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–243, 1984.

[49] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[50] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[51] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust asr," in *INTERSPEECH*, 2012.

[52] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *LVA/ICA*, 2015.

[53] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust asr," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4085–4088, 2012.

[54] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649, IEEE, 2013.

[55] J. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, 2016.

[56] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 787–796, 2018.

[57] Z. qiu Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 686–690, 2018.

[58] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *ArXiv*, vol. abs/1803.01271, 2018.

[59] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *ArXiv*, vol. abs/2002.08933, 2020.

[60] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *ArXiv*, vol. abs/1609.03499, 2016.

[61] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6875–6879, 2019.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[63] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2018.

[64] K. Paliwal, K. Wójcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, pp. 465–494, 2011.

[65] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR*, 2018.

[66] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5069–5073, 2018.

[67] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain," in *INTERSPEECH*, 2018.

[68] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6364–6368, 2020.

[69] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *ICSLP*, 1992.

[70] R. Rehr and T. Gerkmann, "An analysis of noise-aware features in combination with the size and diversity of training data for dnn-based speech enhancement," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 601–605, 2019.

[71] J. L. Roux, S. Wisdom, H. Erdogan, and J. Hershey, "Sdr ? half-baked or well done?," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.

[72] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[73] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017.

[74] M. A. Smith and A. Kohn, "Spatial and temporal scales of neuronal correlation in primary visual cortex," *The Journal of Neuroscience*, vol. 28, pp. 12591 – 12603, 2008.

[75] B. J. Hansen, M. Chelaru, and V. Dragoi, "Correlated variability in laminar cortical circuits," *Neuron*, vol. 76, pp. 590–602, 2012.