Artificial Spectro-Temporal Receptive Fields Evoked by Long Speech Signals

Ondrej Lassak[1], Hans-Heinrich Bothe[2]

[1] *Centre for Applied Hearing Research, Technical University of Denmark, Email: s021548@student.dtu.dk*
[2] *Centre for Applied Hearing Research, Technical University of Denmark, Email: hhb@oersted.dtu.dk*

Introduction

The performance of technical sound or speech processing systems relies strongly on the acoustic or auditory features employed. Our paper proposes the study of biological examples in order to find suitable types of auditory features, since there is a relationship between the statistics of natural stimuli, the properties of the sensory system, and optimal signal processing paradigms.

Neurons in the primary auditory cortex A1 unitize multidimensional receptive activation fields, which respond to specific frequency bands and timing patterns of the input signal. Such ‚spectro-temporal receptive fields' (STRF) can experimentally be determined by reverse correlation methods if the activation or output signal patterns of the neurons are known. This seems impossible for neurons in the human auditory cortex; Thus, only simulations can be done, which show that an evolution of respective banks of STRF is possible or convergent and useful from the perspective of ‚optimal' information processing. In [1], an STRF simulation by Gabor functions or sigma-pi artificial neurons was used to find robust cues for traditional sound processing algorithms. An alternative STRF-based sound processing scheme can be found in [2].

In our paper, we propose to apply independent component analysis (ICA) including a sparse matrix criterion to emulate the evolution of sets of artificial STRF without fitting specific mathematical functions. Due to the sparseness criterion, only few STRF are necessary to represent the input signal at a time, producing an efficient coding procedure and thus, valuable input variables for further processing.

The acoustic speech signal is pre-processed in a biologically plausible form, arriving in the A1 as a modified spectrogram. A bank of STRF acts on that spectrogram, producing temporal courses of 'micro-features' that can be used in further sound or speech processing steps. A typical application of STRF-based features to define speech intelligibility is presented in [3].

Early stages of the auditory pathway

Sound is received by the outer ear before transmitted through the middle ear to the inner ear, where the basilar membrane and the inner hair cells as receptor cells are situated. As the sound waves propagate on the basilar membrane, the upper parts of the inner hair cells, the stereocilia, are sheared back and forth, causing electronic hyperpolarization across the cell membrane and thus, pulse-frequency modulated output signals of the cells. The middle ear can be seen as an impedance converter acting across a large frequency range. The basilar membrane is modelled as a filter bank with a very large number of bandpass filters working on a logarithmic frequency scale.

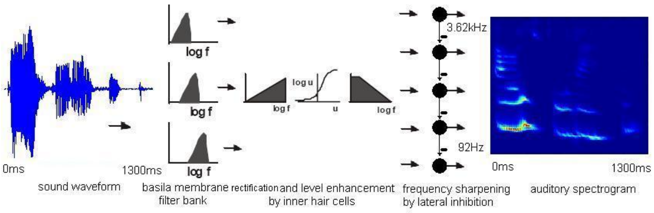
0ms
1300ms
sound waveform
log f
log f
log f
basila membrane filter bank
log f
log u
u
log f
rectification and level enhancement by inner hair cells
3.62kHz
92Hz
frequency sharpening by lateral inhibition
0ms
1300ms
auditory spectrogram

Figure 1: Simplified signal processing steps in the early stages of the auditory pathway, producing an auditory spectrogram representation (adapted from [2]). Rectification and short-time integration after the lateral inhibition is not shown. The resulting auditory spectrogram is the input to STRF-based signal processing in A1.

The information further up in the auditory pathway is pulse-frequency coded. The membrane potentials of the inner hair cells are locked to mechanical oscillations of the basilar membrane. They connect to neurons in the cochlea nucleus, further to the olivary complex, the inferior colliculus, the medial geniculate of the thalamus and finally, to the auditory cortex A1. The exact tasks performed in the single stages are widely unknown – though some properties can be estimated and modelled from neuro-physiological in vitro measurements. A scheme of the signal processing steps in the early stages of the auditory pathway is shown in figure 1. The signal at the input of the inner hair cells can be visualized in form of a spectrogram. This is shaped in the inner hair cells by a high-pass filter due to the fluid-cilia coupling, a rectifying saturation function due the ion channels in the cell membrane and synchronization properties, and a subsequent low-pass filter due to other biochemical properties. Further stages are emulated by a lateral inhibitory neural network that yields a frequency-dependent contrast enhancement of the spectrogram. The auditory spectrogram thus employs a significant logarithmic contrast enhancement with frequency compression.

The neurons in A1 are believed to operate on these modified spectrograms employing individual, localized, frequency and time selective receptive fields, the STRF. Optimum stimulation of a neuron occurs when the auditory spectrogram changes in frequency and time according to its STRF. The matrix of STRF is sparse-coded in the sense that only few neurons are necessary to represent the spectrogram.

Statistical models and ICA

Starting from the auditory spectrogram, we generate biologically-inspired artificial STRF, which can be used as features in technical sound and speech processing tasks. We employ independent component analysis (ICA) that leads to a sparse-coded representation of the signal.

ICA is a matrix factorization method factorizing the probability to generate observed data $\mathbf{x_i}$ with the help of a linear mixing of independent sources $\mathbf{s_i}$ [4]. The observed data are short temporal excerpts of the auditory spectrogram, the sources are the requested STRF, and the mixing matrix **A** determines the sparseness of the transformation. The mixing matrix as well as the sources are a priori unknown. The remixing problem is solved by considering statistical properties of the independent sources $\mathbf{s_i}$. The basic problem is to find a transformation, i.e. a mixing matrix **A** that generates m observed data vectors $\mathbf{x_i}$ (sampled spectro-gram vectors) from unknown but statistically independent sources $\mathbf{s_i}$. The observed data vectors $\mathbf{x_i}$ are arranged in columns to form the data matrix **X**, the unknown sources $\mathbf{s_i}$ to form the matrix **S**, and the unknown mixing features $\mathbf{a_k}$ the mixing matrix **A**.

$$X = A \cdot S \tag{1}$$

ICA now seeks to find a mixing matrix **A** that optimizes the statistical independence of **S** under the constraints of the sparseness of the coefficients in **A**. The optimization process uses statistical or entropy measures related to the kurtosis, mutual information, or negentropy. Eventually, ICA produces a set of non-orthogonal features and as such, it creates a set of sparsely distributed variables which effectively encode the input data of the auditory spectrogram.

With respect to the optimization method used, different ICS paradigms can be created. We employed several ICA paradigms and, in order to be independent on specific speech utterances, a long speech signal as the input. The analysis thus results in different sets of STRF, depending on the sparseness constraints employed. The method that produced STRF most similar to biological examples is TICA [4].

Text corpus

A nine minutes long speech utterance produced by a female speaker was used as the text corpus [5]. The speaker reads parts of the first chapter of a book in Czech language with normal speed and articulation.

Simulation with TICA

TICA has been used to model complex cells in the primary visual cortex V1, but it can also be employed to generate artificial auditory STRF. We used typical durations T of the excerpts of the spectrograms between 25-100 ms and a maximum frequency f_{max}=3.6 kHz. In figure 2, the first 80 resulting STRF of an exemplary set of 1024 STRF are shown. The duration of the training data was 60 ms. Black colour indicates minimum (negative) amplitude, grey the zero level, and white maximum amplitude. The level resolution is eight bits. The localized nature of the resulting STRF can nicely be seen. They consist of non-separable spectral and temporal terms and generally look like chirp waves pointing up- or downwards. The sparseness of the STRF can be estimated by the total relative time during which the single STRF are used during the complete input signal; this is for any STRF approximately 2%. The average overlap neural activation per training interval is approximately 5%. The STRFs can now be used as independent filter functions, each of which creating a course of an auditory feature for further signal processing. The resulting data streams can, for example, be used for signal prediction, sound or speech classification, recognition, or identification.

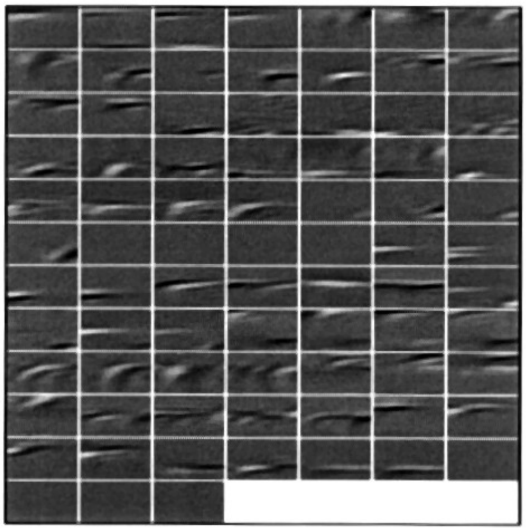

Figure 2: Set of 80 trained STRF. The zero level is shown in grey; white corresponds to positive, black to negative levels. The length of each STRF is 60 ms, the frequency range spans logarithmically between 90Hz and 3.6kHz. The frequency axis points downwards.

Conclusions

Topographical ICA can successfully be employed to create sparsely distributed sets of local STRF that show similarity to measured biological STRF of A1 cells in animals.

References

[1] Kleinschmidt, M., Tchorz, J., Kollmeier, B., Combining speech enhancement and auditory feature extraction for robust speech recognition, Speech Communication, **34,** 1-2, (2001), 75-91.

[2] Wang, K., Shamma, S., Representation of spectral profiles in primary auditory cortex, IEEE Trans. Speech and Audio Proc. **3** (1995), 382-395.

[3] Elhilali M., Chi T. and Shamma S., Intelligibility and the spectrotemporal representation of speech in the auditory cortex, Speech Communication **41** (2003), 331-348.

[4] Hyvärinen, A., Hoyer, P.O., Inki, M., Topographic independent component analysis, Neural Computation **13.7** (2001), 1527-1558.

[5] Hana Maciuchova reading Alois Jirásek, Die Sagen Böhmens - Staré povìsti èeské.
URL: http://www.geocities.com/phonik2/maciuchova.html.