Modeling of Polish Intonation for Statistical-Parametric Speech Synthesis

Tomasz Kuczmarski

Prof. zw. Dr hab. Inż. Grażyna Demenko

Supervisor

Adam Mickiewicz University

Faculty of Modern Languages and Literature
Institute of Ethnolinguistics

May 18, 2022

Intonation

Definition

"The tonal structure of speech expressed by the melody produced by our larynx. It has a phonetic aspect, *the fundamental frequency (F0)*, and a grammatical (phonological) aspect." *(Féry 2016)*

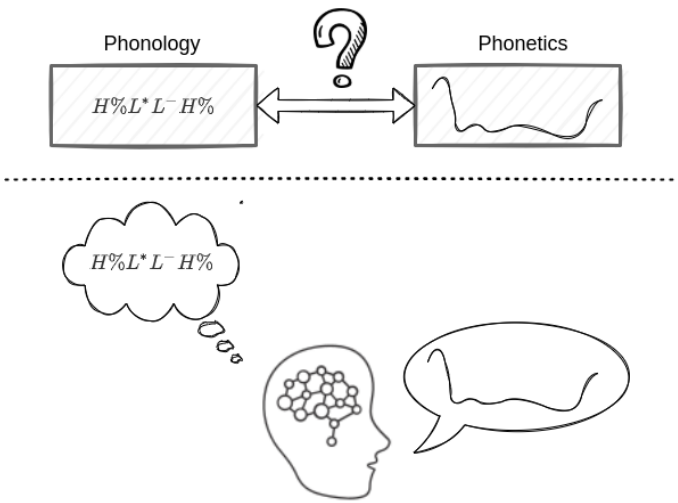
Intonation

Definition

"The tonal structure of speech expressed by the melody produced by our larynx. It has a phonetic aspect, *the fundamental frequency (F0),* and a grammatical (phonological) aspect." *(Féry 2016)*

All definitions of intonation "are epistemological definitions,i.e., not a priori programmatic definitions, but a posteriori statements of a practice and methodology." (Rossi 2000)

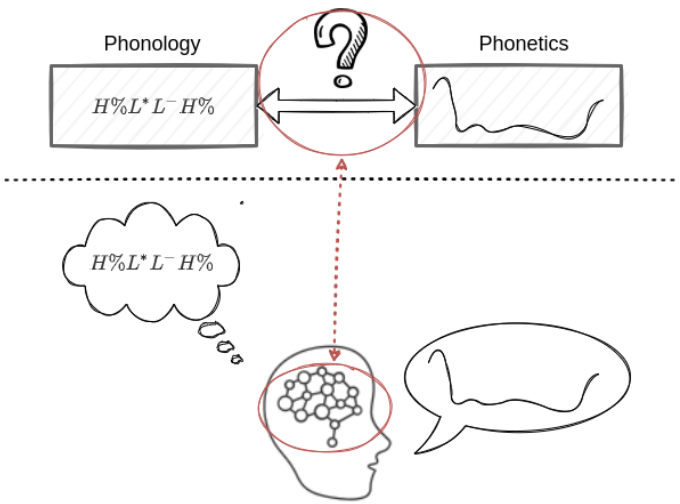
Motivation



Motivation

- ✓ Epistemological definition of intonation.
- ✓ Dualistic gap between phonology and phonetics.
- ✓ Unification within a broader metatheory.
- ✓ Unknown nature of the mappings between mental categories and continuous contours of $F0$.
- ✓ Need for a physicalist (neurobiological) model.
- ✓ Modern statstical-parametric speech synthesis provides a framework for experimentation and evaluation of such models.

Motivation



Motivation

- ✓ Epistemological definition of intonation.
- ✓ Dualistic gap between phonology and phonetics.
- ✓ Unification within a broader metatheory.
- ✓ Unknown nature of the mappings between mental categories and continuous contours of $F0$.
- ✓ Need for a physicalist (neurobiological) model.
- ✓ Modern statstical-parametric speech synthesis provides a framework for experimentation and evaluation of such models.

Objectives

1. Build a robust biologically-inspired neural model of the probabilistic mapping between discrete low-level linguistic features of an utterance and its intonation contours ($F0$ values).
2. Build a state-of-the-art neural source-filter resynthesis framework for Polish read speech.
3. Deploy the intonation model within the resynthesis framework and measure the perceived naturalness of the output intonation contours, and to
4. operationalize the results of these measurements as an indicator of the model's robustness.
5. Develop a method to explain the relevance of the individual input linguistic features for the produced intonation contour.
6. Analyze how specific linguistic features contribute to the $F0$ contours of an utterance.

Objectives

1. Build a robust biologically-inspired neural model of the probabilistic mapping between discrete low-level linguistic features of an utterance and its intonation contours ($F0$ values).
2. Build a state-of-the-art neural source-filter resynthesis framework for Polish read speech.
3. Deploy the intonation model within the resynthesis framework and measure the perceived naturalness of the output intonation contours, and to
4. operationalize the results of these measurements as an indicator of the model's robustness.
5. Develop a method to explain the relevance of the individual input linguistic features for the produced intonation contour.
6. Analyze how specific linguistic features contribute to the $F0$ contours of an utterance.

Main Hypotheses

HYPOTHESIS 1: *The continous F_0 contours of an utterance emerge from its discrete linguistic features through a series of successive probabilistic mappings into intermediate latent represenentations.*

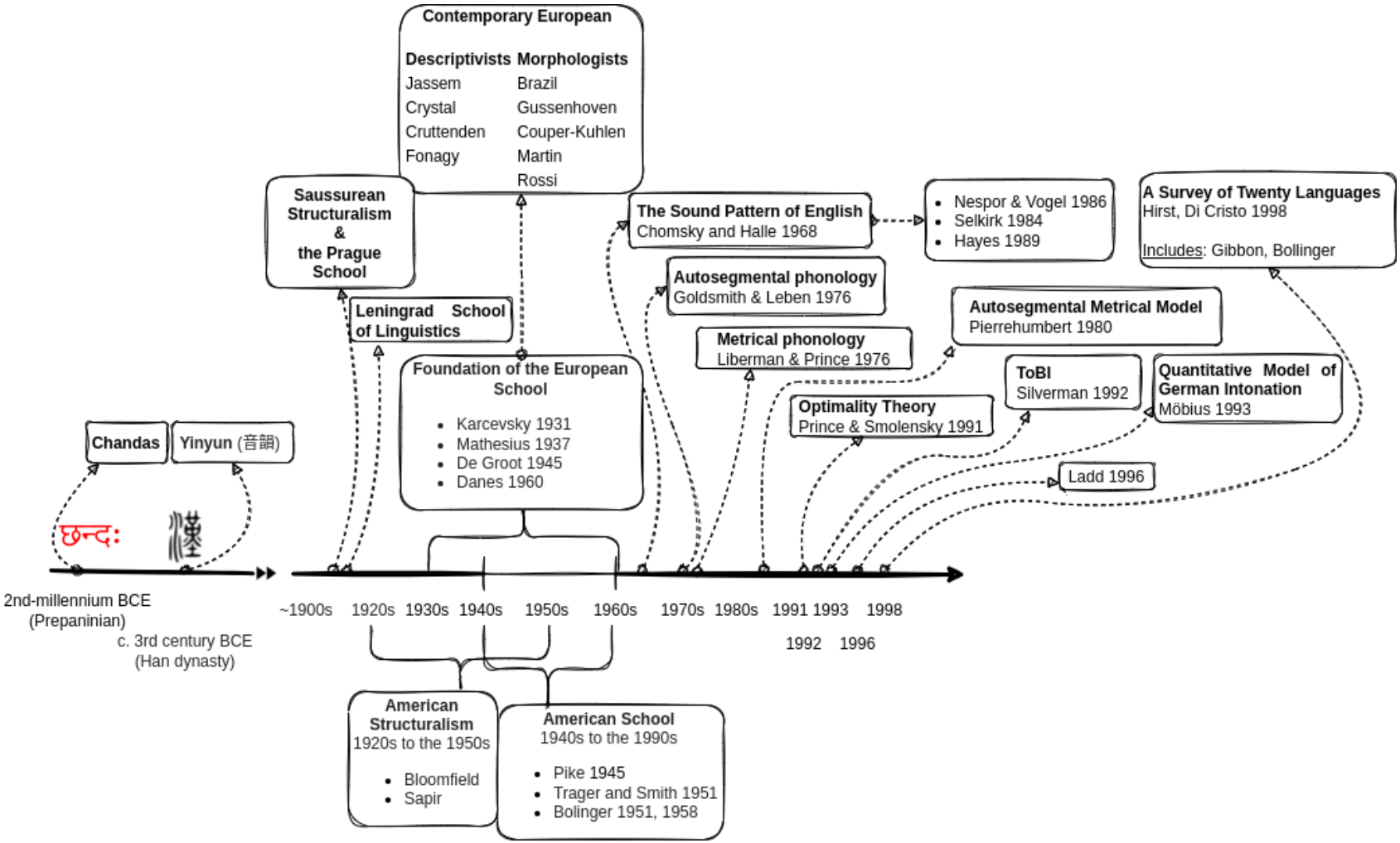
HYPOTHESIS 2: *The biologically-inspired Deep Temporal Convolutional Network can be an effective model of these mappings and hence of Polish neutral read speech intonation in the context of statistical-parametric speech synthesis.*

HYPOTHESIS 3: *The set of shallow linguistic features used in this thesis provides information which is sufficient for synthesis of natural sounding intonation in the context of statistical-parametric speech synthesis.*

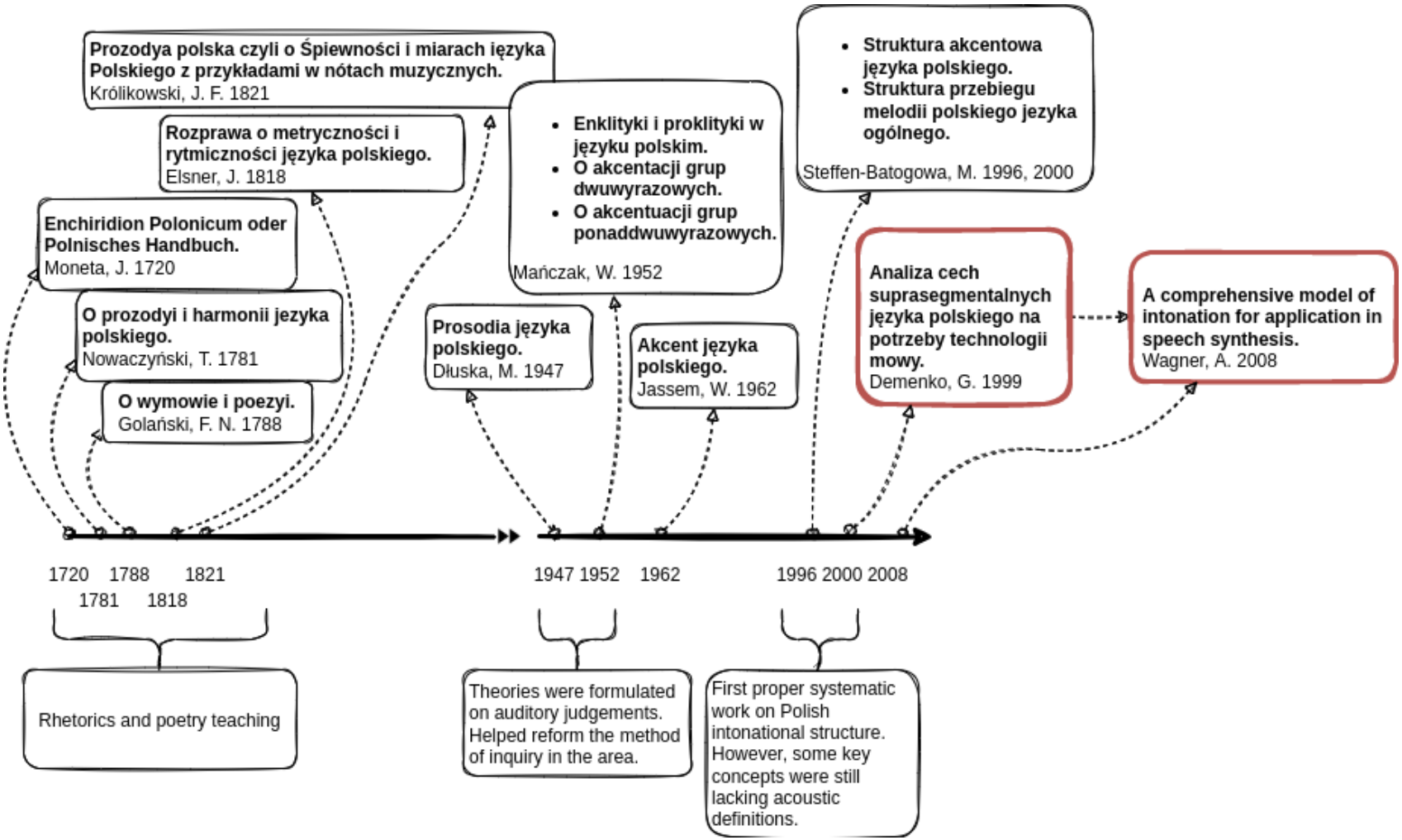
Contributory Methodological Hypothesis

HYPOTHESIS 4 (CONTRIBUTORY METHODOLOGICAL): *A Deep Temporal Convolutional Network can become an explanatory scientific model of mappings between linguistics features and the intonation of an utterance.*

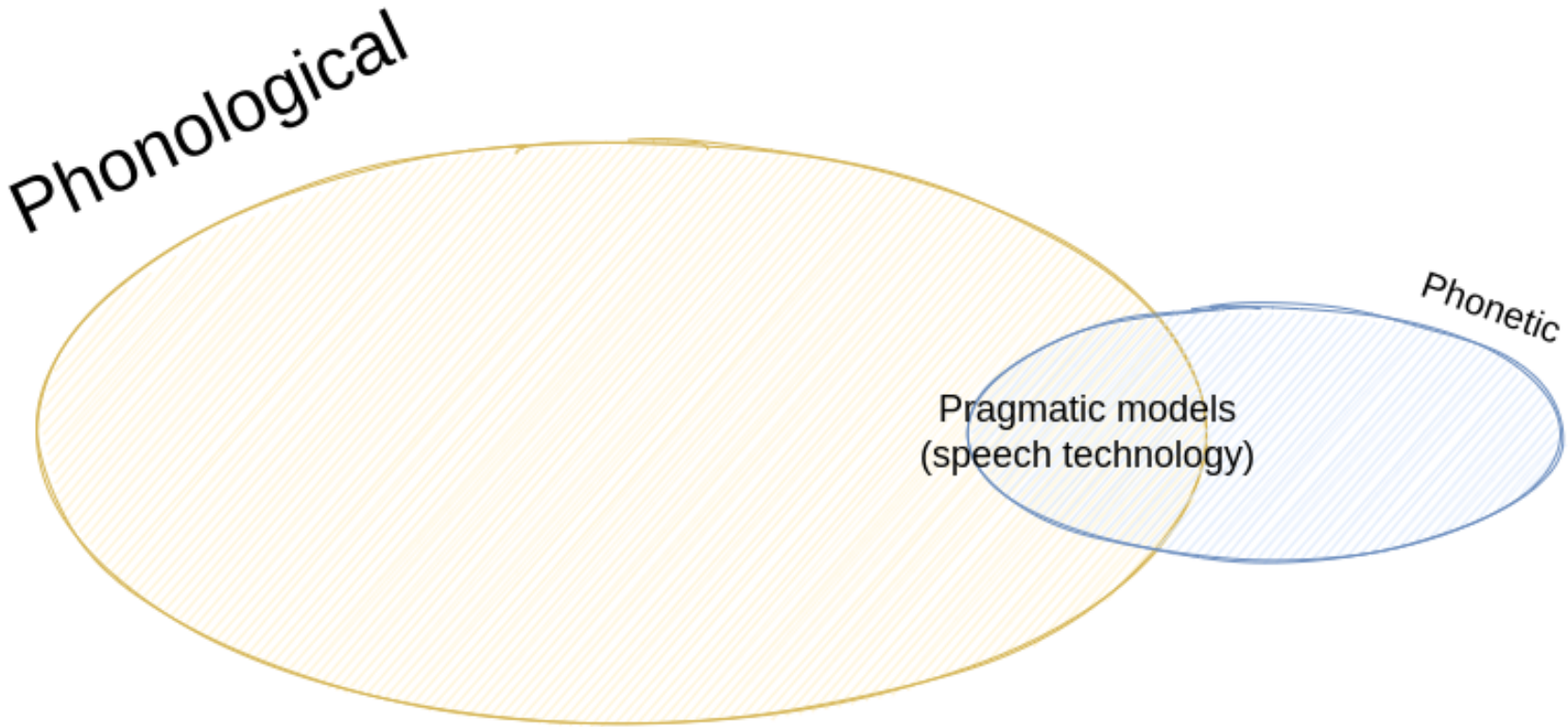
Background


Contemporary European
Descriptivists Morphologists
Jassem Brazil
Crystal Gussenhoven
Cruttenden Couper-Kuhlen
Fonagy Martin
Rossi
Saussurean Structuralism & the Prague School
Leningrad School of Linguistics
Foundation of the European School
Karcevsky 1931
Mathesius 1937
De Groot 1945
Danes 1960
The Sound Pattern of English
Chomsky and Halle 1968
Nespor & Vogel 1986
Selkirk 1984
Hayes 1989
A Survey of Twenty Languages
Hirst, Di Cristo 1998
Includes: Gibbon, Bollinger
Autosegmental phonology
Goldsmith & Leben 1976
Autosegmental Metrical Model
Pierrehumbert 1980
Metrical phonology
Liberman & Prince 1976
ToBI
Silverman 1992
Quantitative Model of German Intonation
Möbius 1993
Optimality Theory
Prince & Smolensky 1991
Ladd 1996
Chandas
Yinyun (音韻)
छन्दः
2nd-millennium BCE
(Prepaninian)
c. 3rd century BCE
(Han dynasty)
~1900s 1920s 1930s 1940s 1950s 1960s 1970s 1980s 1991 1993 1998
1992 1996
American Structuralism
1920s to the 1950s
Bloomfield
Sapir
American School
1940s to the 1990s
Pike 1945
Trager and Smith 1951
Bolinger 1951, 1958

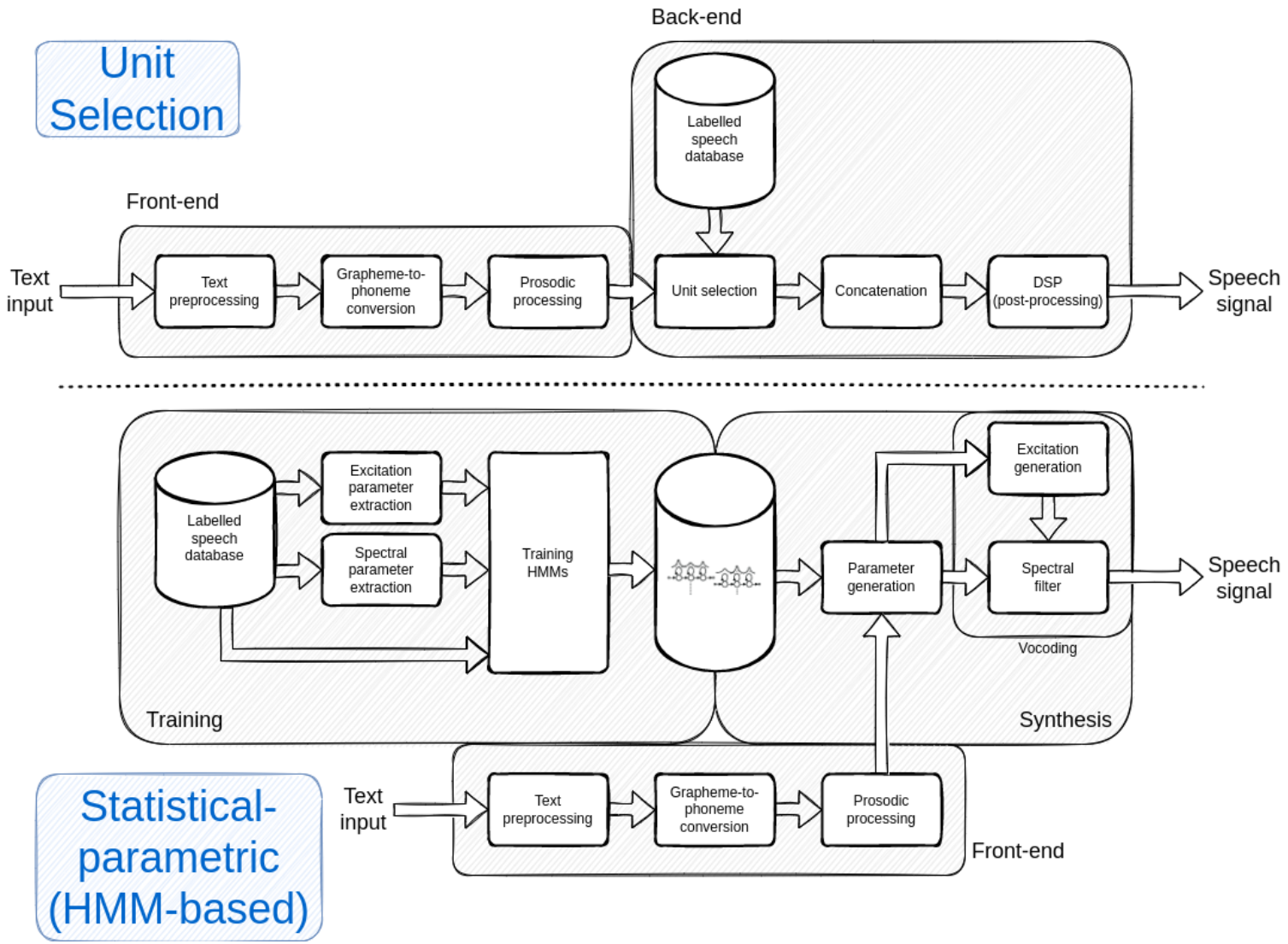
Background


Prozodya polska czyli o Śpiewności i miarach języka Polskiego z przykładami w nótach muzycznych.
Królikowski, J. F. 1821
Rozprawa o metryczności i rytmiczności języka polskiego.
Elsner, J. 1818
Enchiridion Polonicum oder Polnisches Handbuch.
Moneta, J. 1720
O prozodyi i harmonii jezyka polskiego.
Nowaczyński, T. 1781
O wymowie i poezyi.
Golański, F. N. 1788
Enklityki i proklityki w języku polskim.
O akcentacji grup dwuwyrazowych.
O akcentuacji grup ponaddwuwyrazowych.
Mańczak, W. 1952
Prosodia języka polskiego.
Dłuska, M. 1947
Akcent języka polskiego.
Jassem, W. 1962
Struktura akcentowa języka polskiego.
Struktura przebiegu melodii polskiego jezyka ogólnego.
Steffen-Batogowa, M. 1996, 2000
Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy.
Demenko, G. 1999
A comprehensive model of intonation for application in speech synthesis.
Wagner, A. 2008
1720 1781 1788 1818 1821
1947 1952 1962
1996 2000 2008
Rhetorics and poetry teaching
Theories were formulated on auditory judgements. Helped reform the method of inquiry in the area.
First proper systematic work on Polish intonational structure. However, some key concepts were still lacking acoustic definitions.

Intonation models

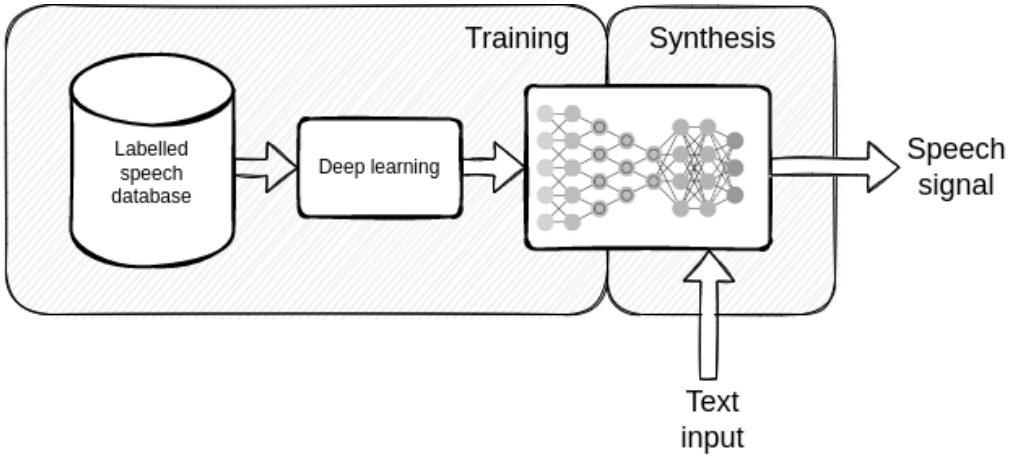


Speech synthesis



Speech synthesis - Wavenet

Wavenet



Wavenet belongs to a class of models known as Convolutional Neural Networks (CNNs).

Speech synthesis - Wavenet

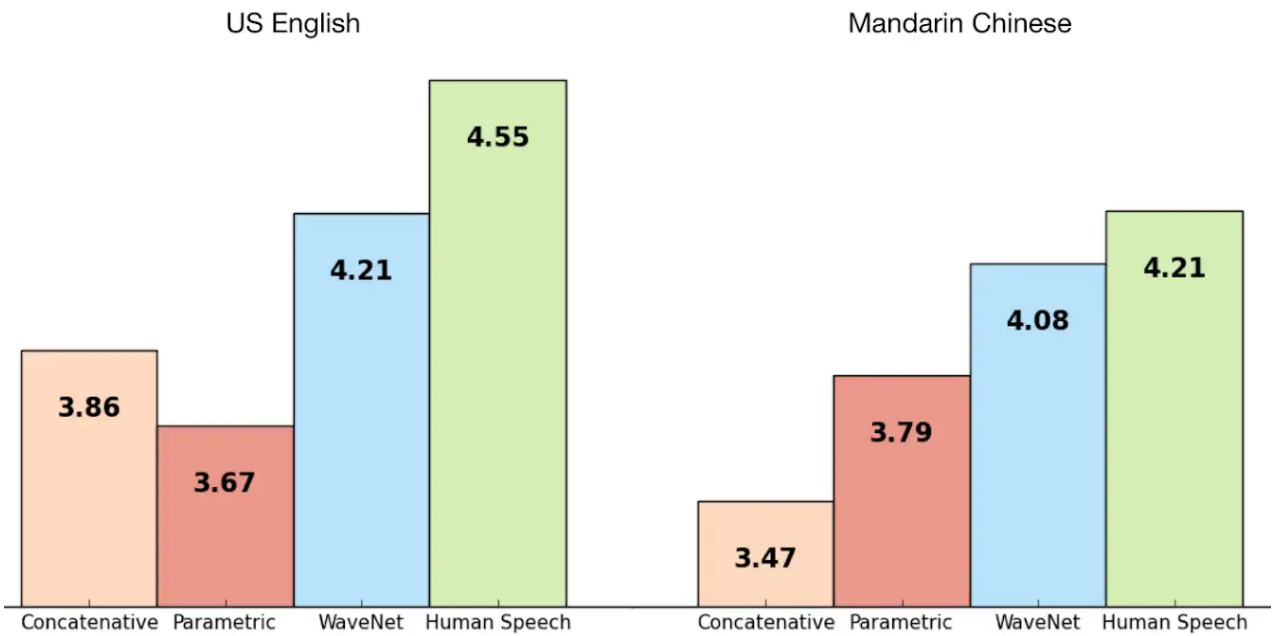


Figure: Google WaveNet evaluation results as compared with Google's best concatenative and parametric systems. (from van den Oord 2016).

The idea behind Wavenet is based on a PixelCNN Image Generator also by developed by van den Oord (2016) at Google's DeepMind.

Neurobiological foundations

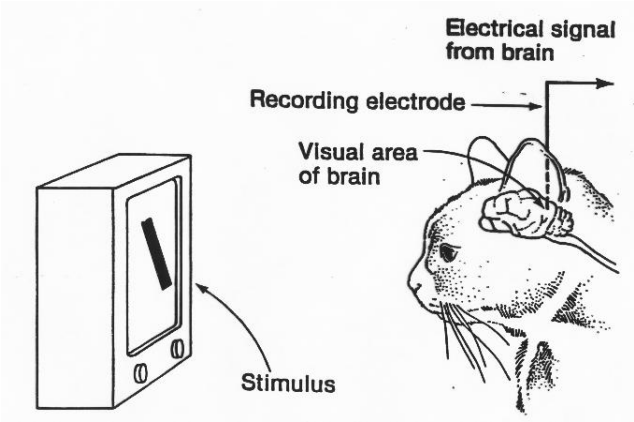


Figure: Famous Hubel and Wiesel cat experiment. (adopted from Hubel and Wiesel 1959).

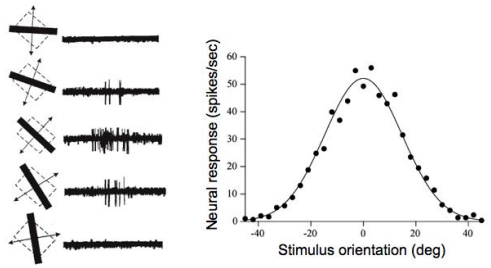


Figure: Neural response of simple cells. (adopted from Hubel and Wiesel 1968).

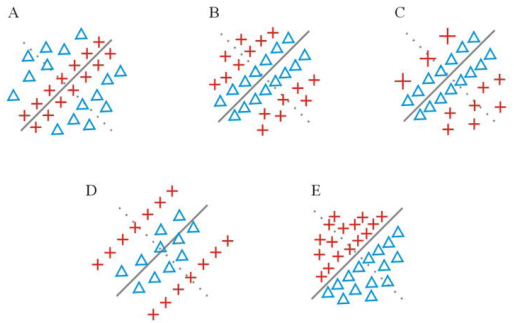


Figure: Simple receptive fields. (adopted from Hubel and Wiesel 1962).

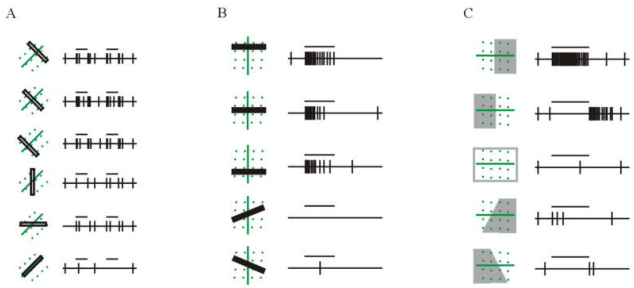


Figure: Three different types of complex receptive fields. (adopted from Hubel and Wiesel 1962).

Neurobiological foundations

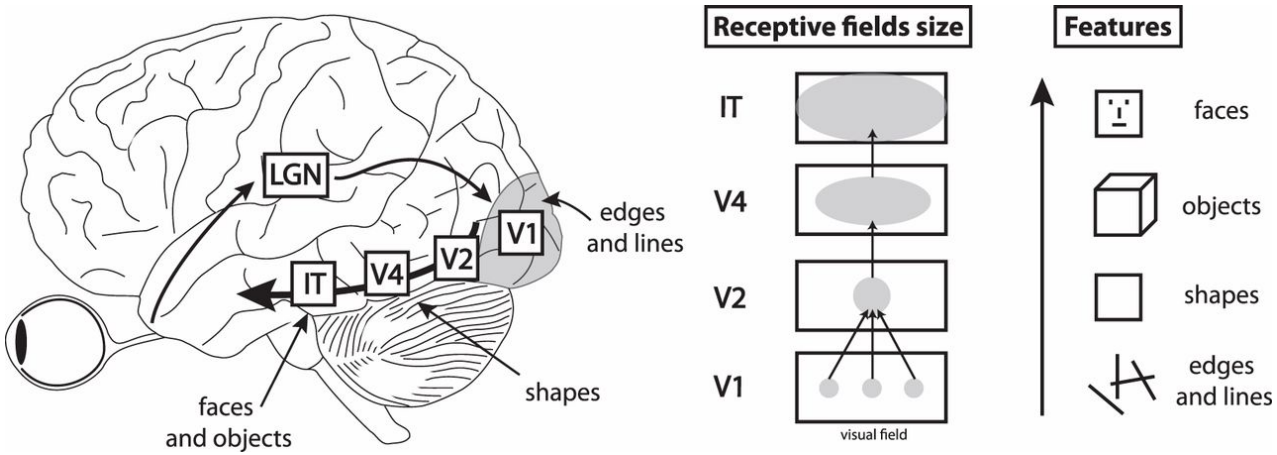


Figure: Hierarchical, feedforward visual processing in human brain. (Adopted from Manassi et al. 2013)

Neurobiological foundations

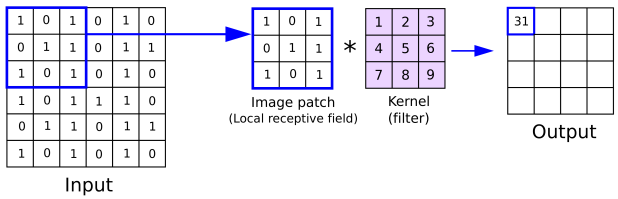


Figure: Example of a 2-dimensional matrix convolution.

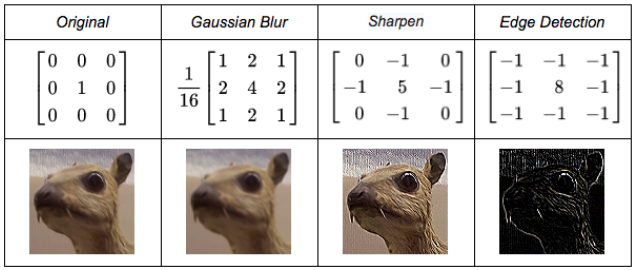


Figure: Examples of convoluting and image with different convolution kernels. (Adopted from the Wikipedia).

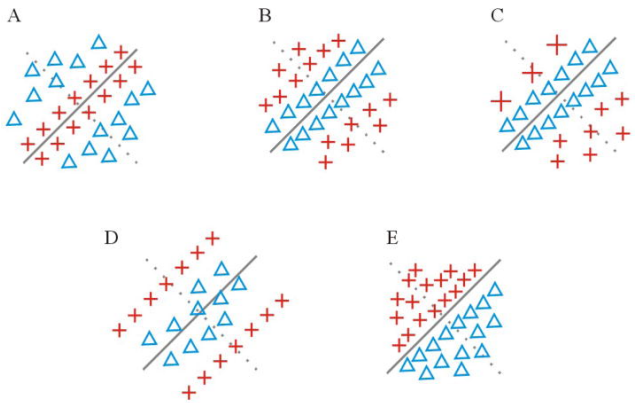


Figure: Simple receptive fields. (adopted from Hubel and Wiesel 1962).

Simple cells perform edge and line detection which can be very effectively approximated with matrix convolution.

Neurobiological foundations

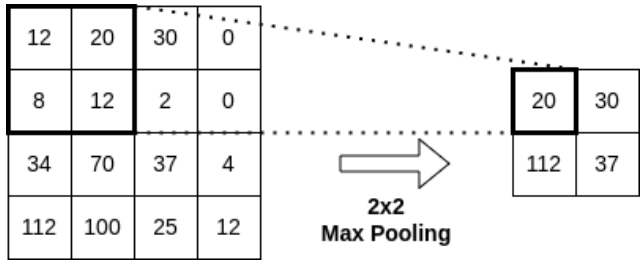


Figure: Example of a 2x2 max pooling matrix operation.

The function of the complex cells can be well approximated by the max pooling operation.

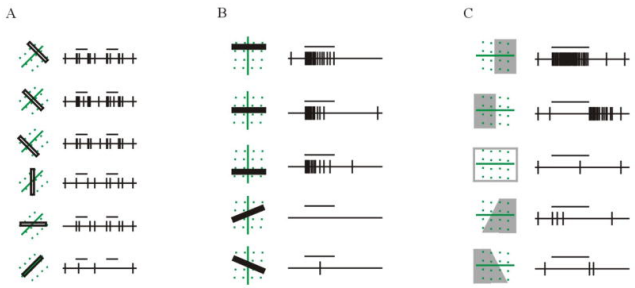


Figure: Three different types of complex receptive fields. (adopted from Hubel and Wiesel 1962).

Neurobiological foundations

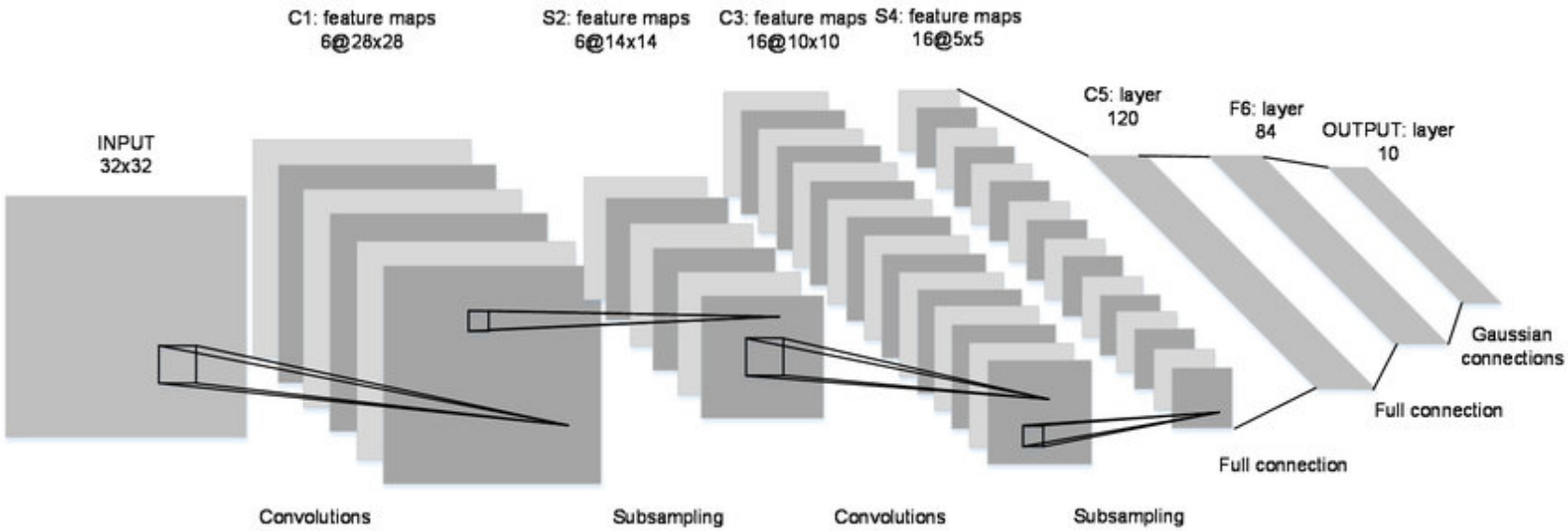
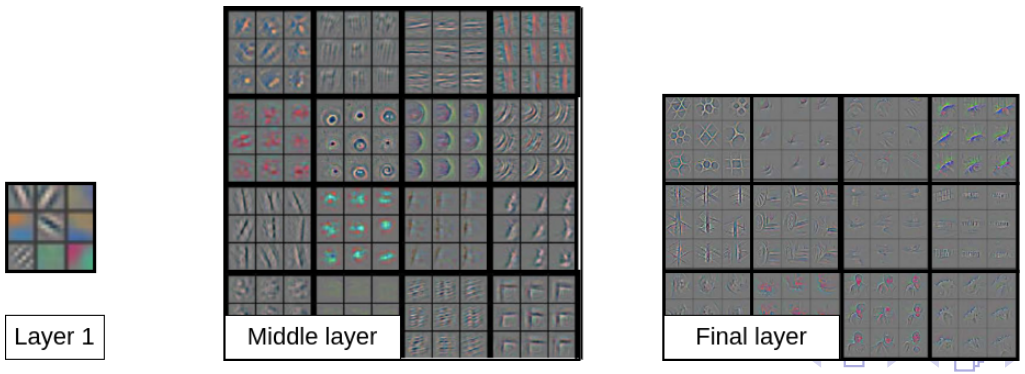


Figure: Image recognition Convolutional Neural Network (LeNet-5). (Adopted from LeCun et al. 1989)



Neural auditory processing

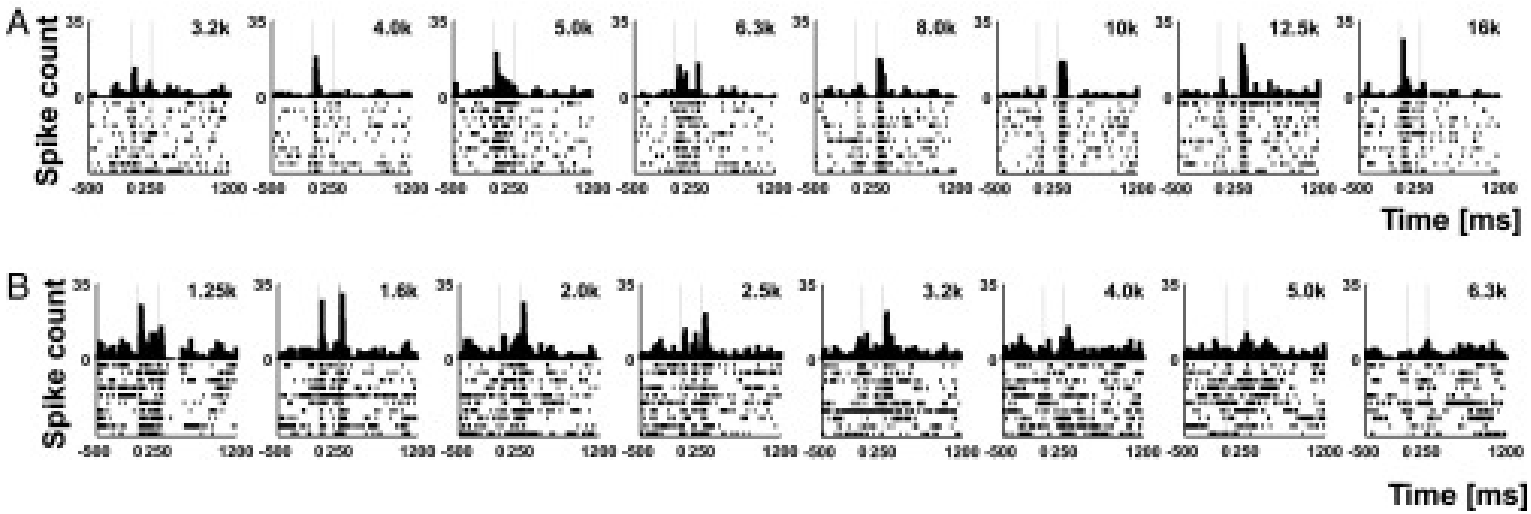


Figure: Responses of single neurons in primary auditory cortex (A1) of rhesus monkeys to band-passed noise (BPN) bursts centered at particular frequencies.. (Adopted from Tian et al. 2013)

Auditory cortex also contains simple cells with two dimensional receptive fields that are similar to those in the visual cortex but operate in the spectrotemporal domain.

Explanatory properties of CNNs

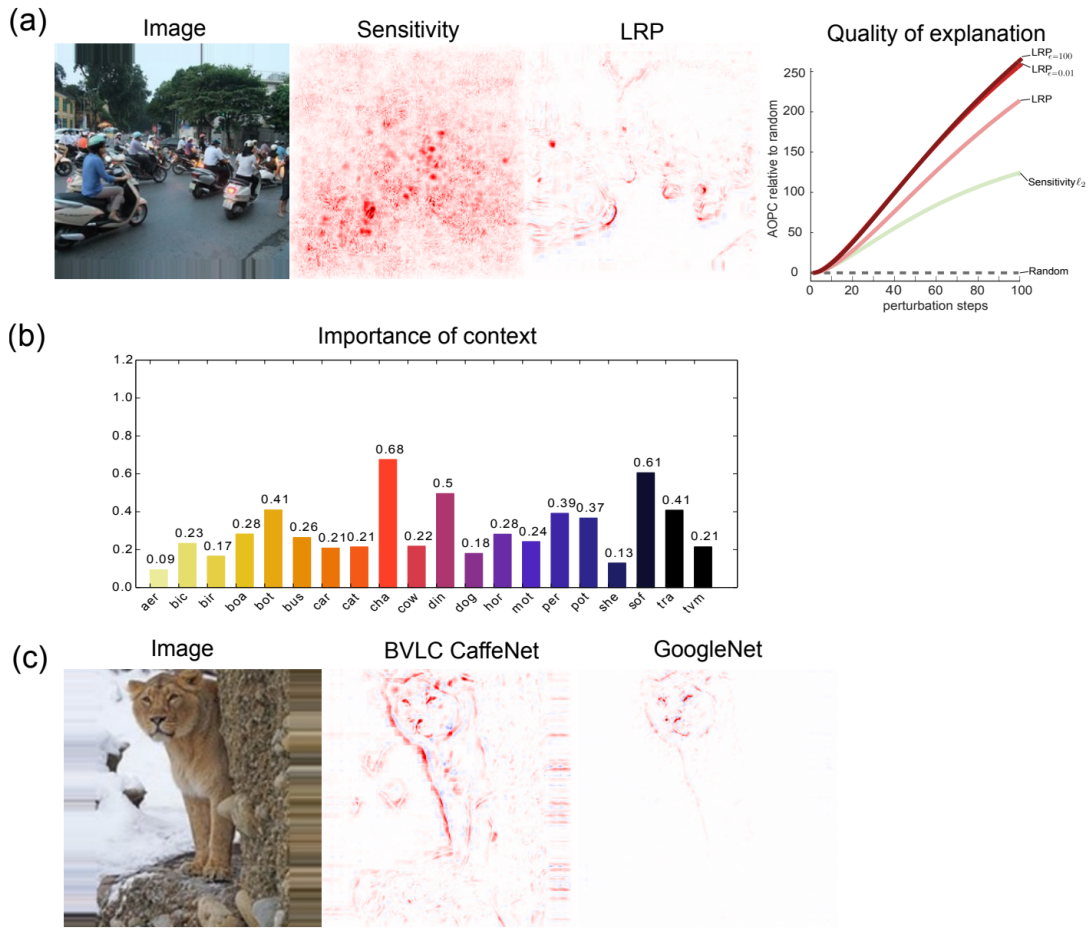


Figure: Results of sensitivity-based and relevance-based explainability methods. (Based on Samek 2016)

Deep Taylor Decomposition

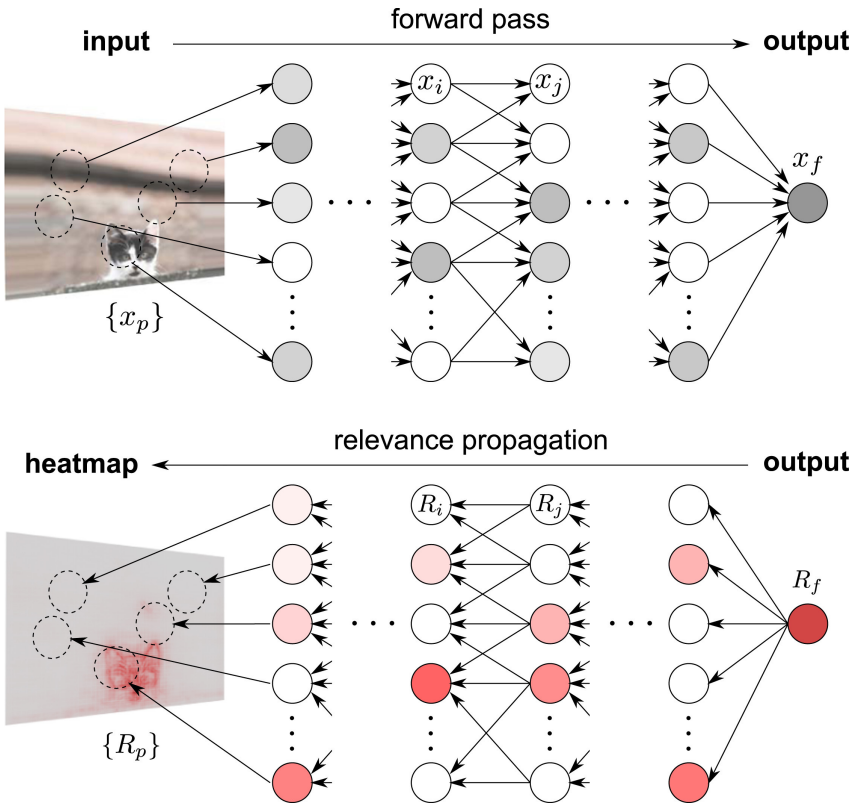


Figure: Computational flow of deep Taylor decomposition. (Adopted from Montavon 2017).

The relevance in this framework can be defined as:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k \qquad (1)$$

Thank you.