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)

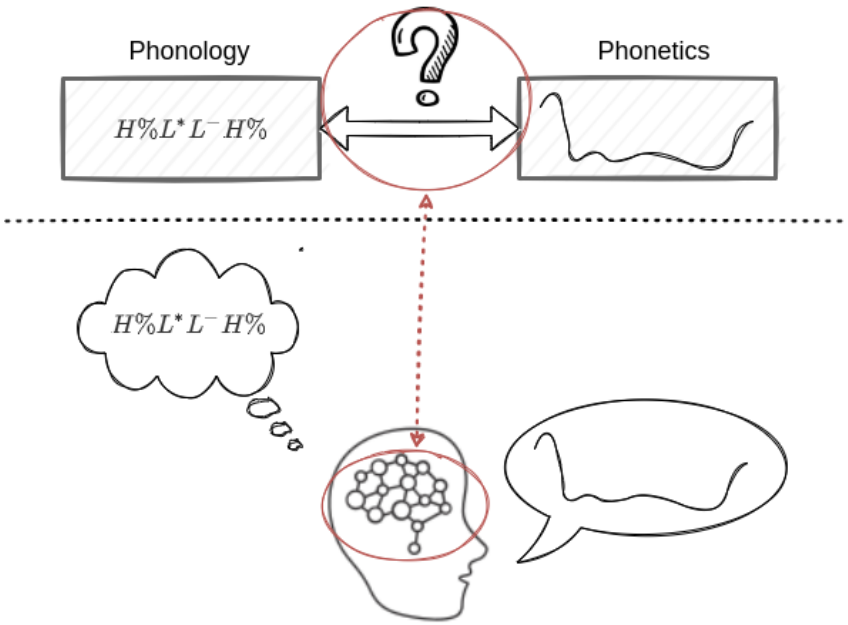
Intonation

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

- ✓ Dualistic gap between phonology and phonetics.
- ✓ Unification within a broader metatheory.
- ✓ Unknown nature of the mappings between mental categories and continuous contours of $F0$.
- ✓ How linguistic features of an utterance influence its $F0$ contours.
- ✓ Need for a physicalist (neurobiological) model.
- ✓ Modern statstical-parametric speech synthesis provides a framework for experimentation and evaluation of such
- ✓ Remarkable properties of biologically-motivated Convolutional Neural Networks.

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.*

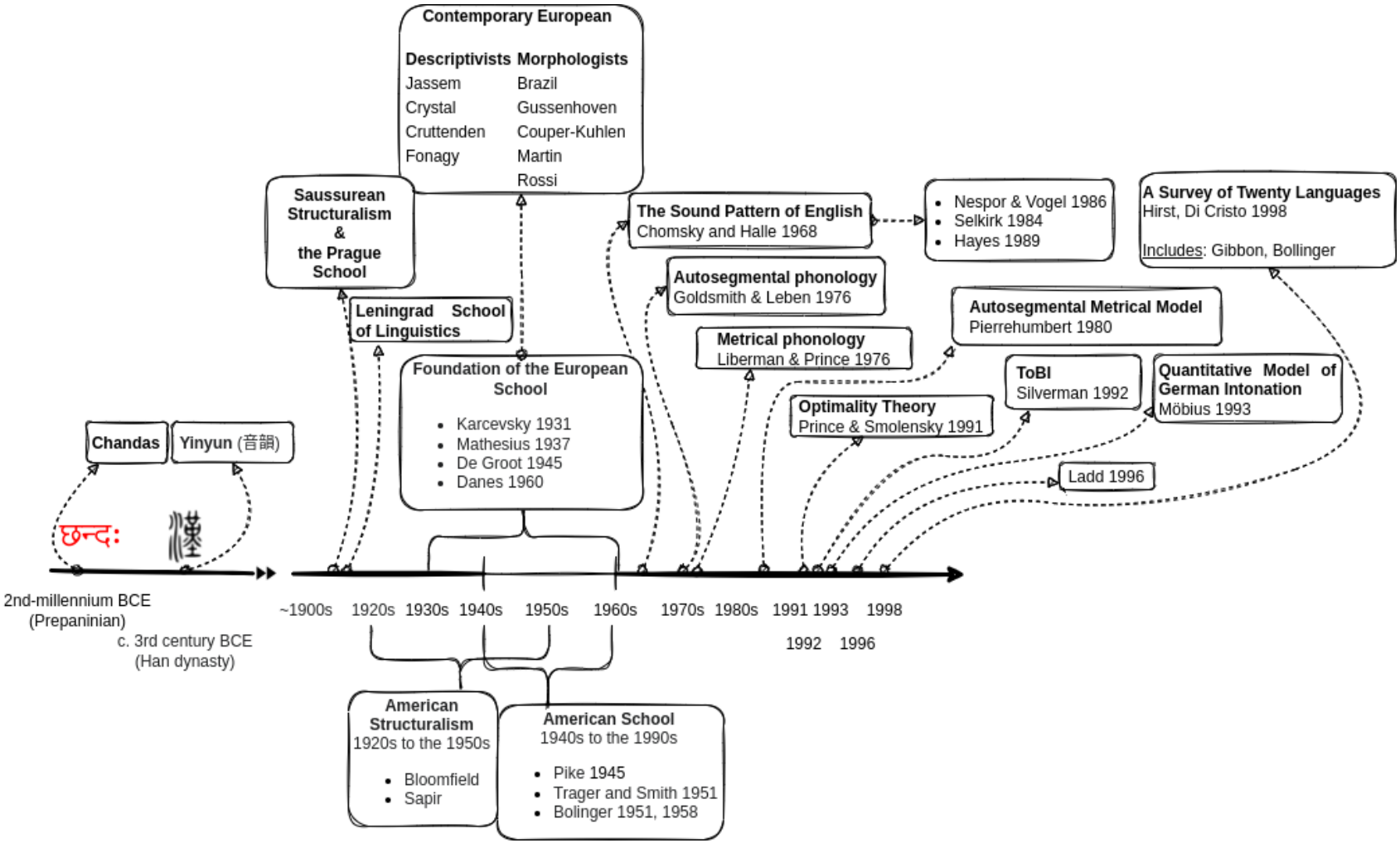
HYPOTHESIS 2: *The biologically-inspired Deep Temporal Convolutional Network can be an effective model of these mappings and hence of Polish neutral read speech intonation in the context of statistical-parametric speech synthesis.*

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.*

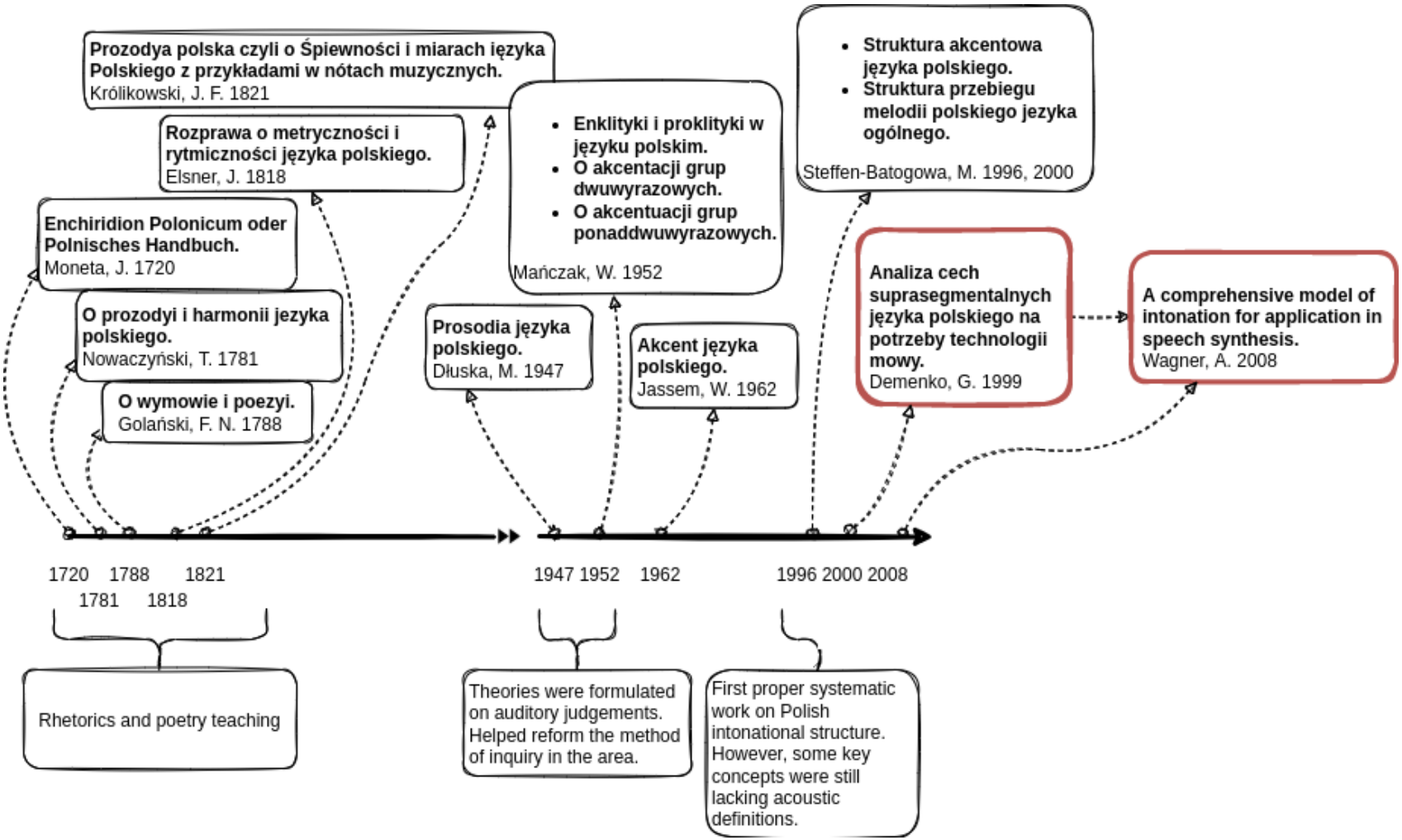
Contributory Methodological Hypothesis

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.*

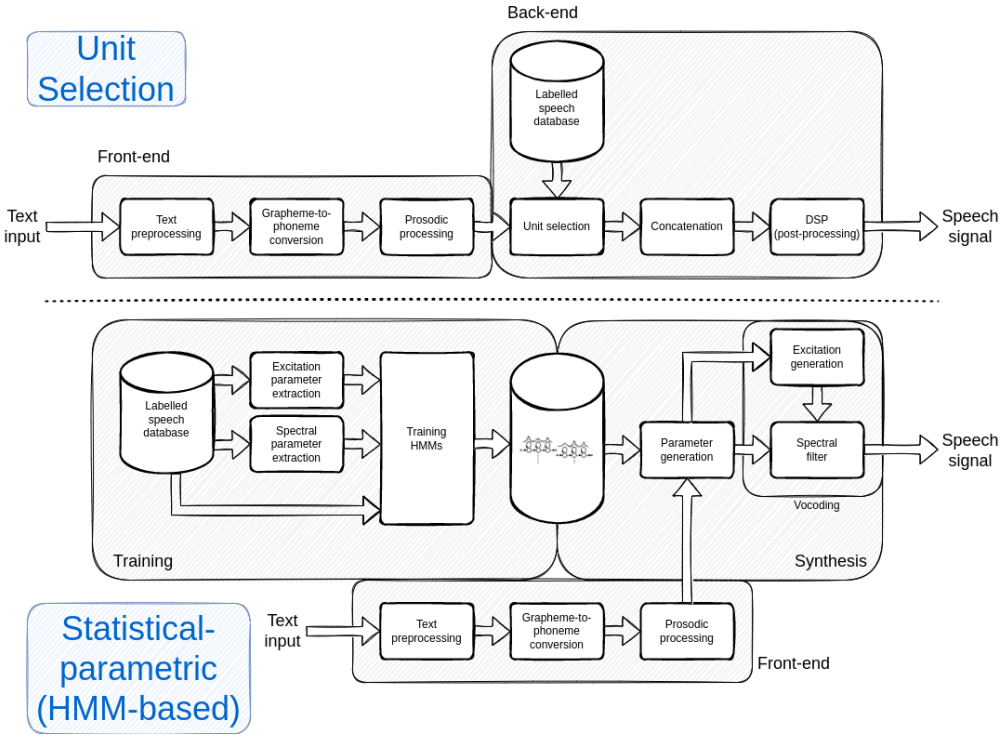
Background

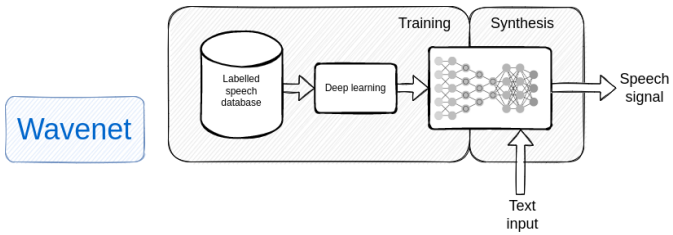

Contemporary European
Descriptivists
Jassem
Crystal
Cruttenden
Fonagy
Morphologists
Brazil
Gussenhoven
Couper-Kuhlen
Martin
Rossi
Saussurean Structuralism & the Prague School
Leningrad School of Linguistics
Foundation of the European School
Karcevsky 1931
Mathesius 1937
De Groot 1945
Danes 1960
The Sound 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
1992
1993
1996
1998
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.

Speech synthesis





Speech synthesis - Wavenet

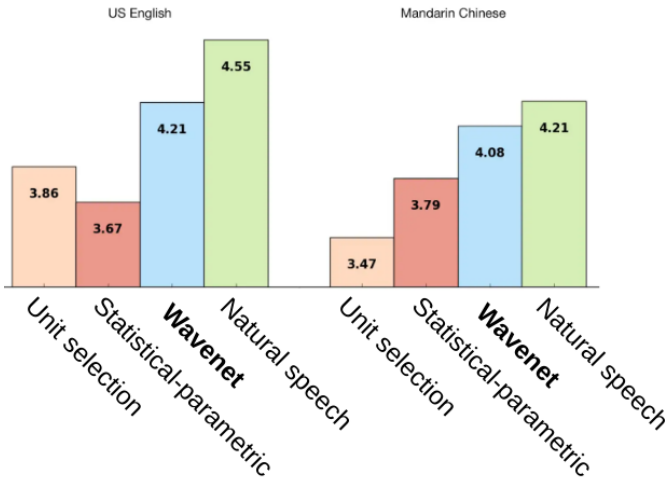
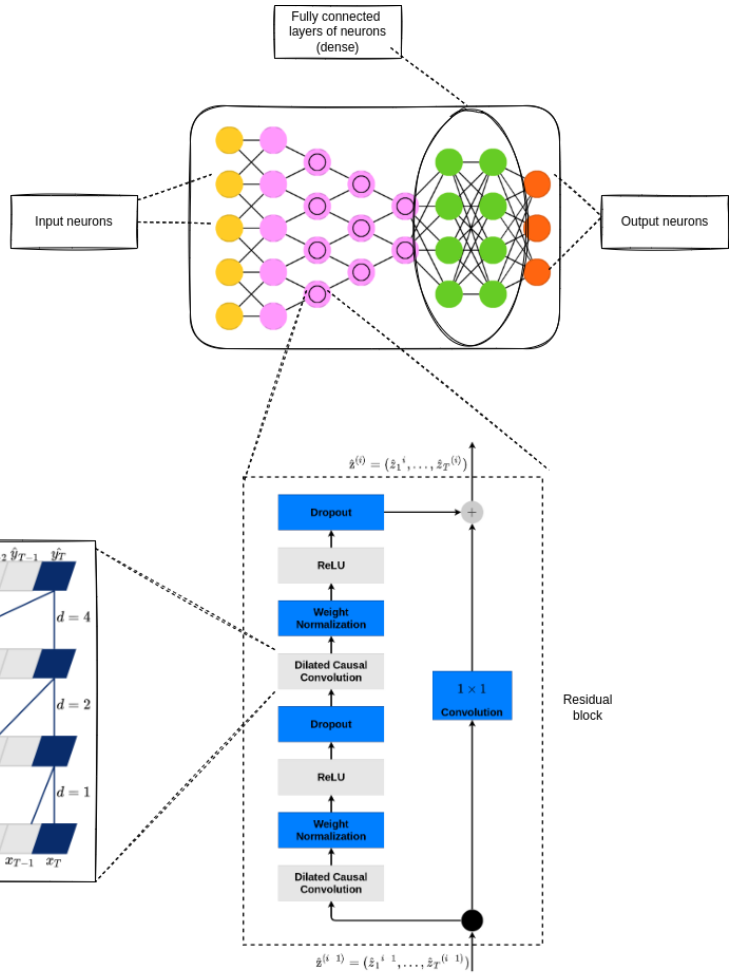


Figure: Google WaveNet evaluation results. (from van den Oord 2016).

Wavenet belongs to a class of models known as **Convolutional Neural Networks** (CNNs) which are used mainly in the area of of image recognition where they excel.



Visual processing in the human brain

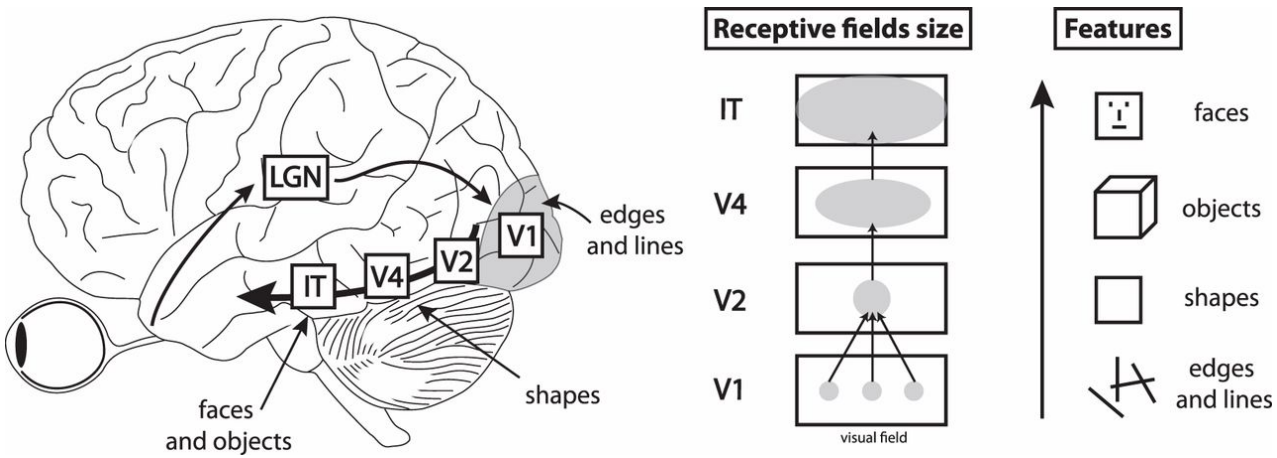


Figure: Hierarchical, feedforward visual processing in human brain. (Adopted from Manassi et al. 2013)

CNN feature maps

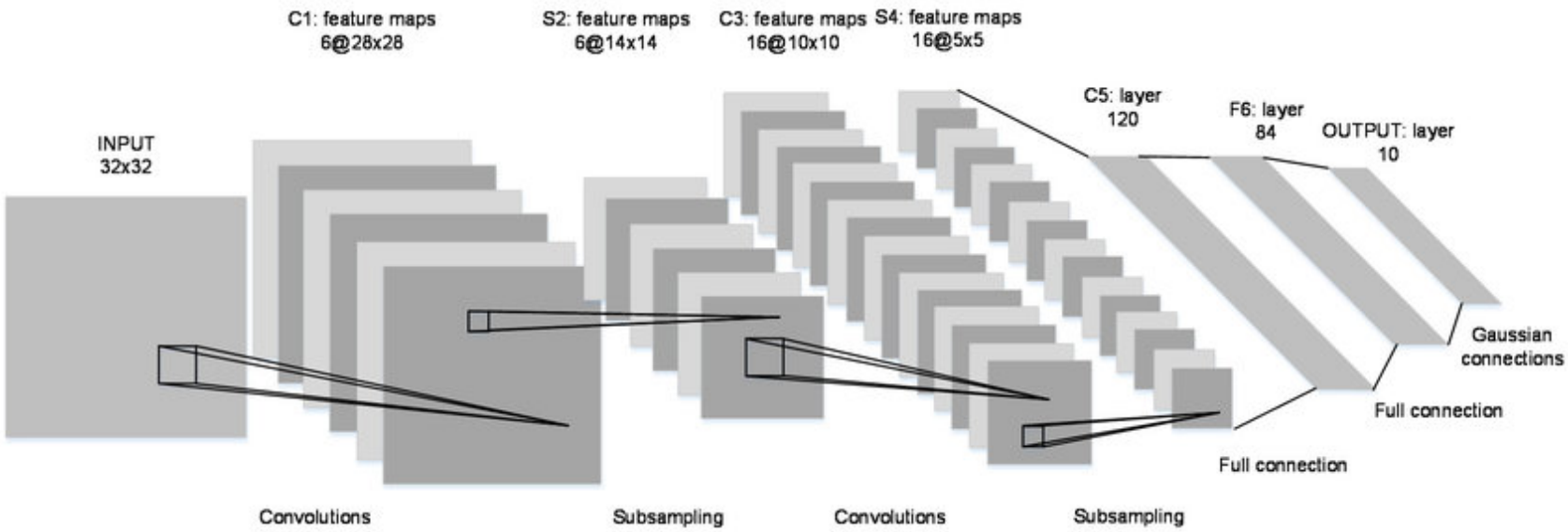
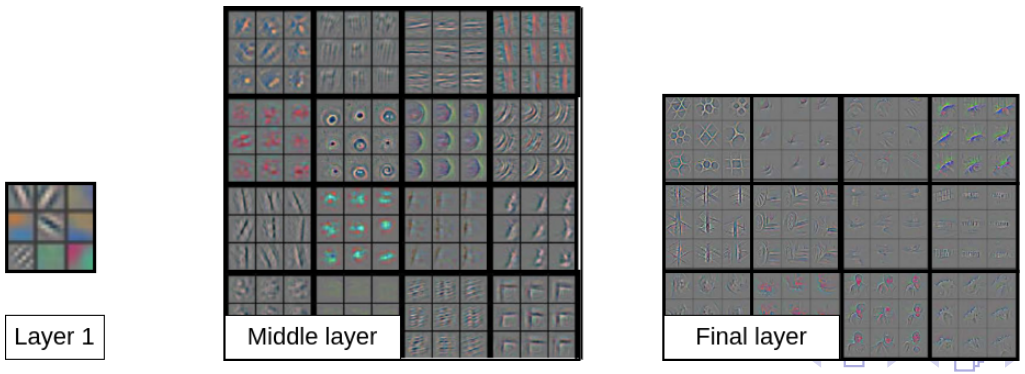


Figure: Image recognition Convolutional Neural Network (LeNet-5). (Adopted from LeCun et al. 1989)



Simple and complex cells in the visual cortex

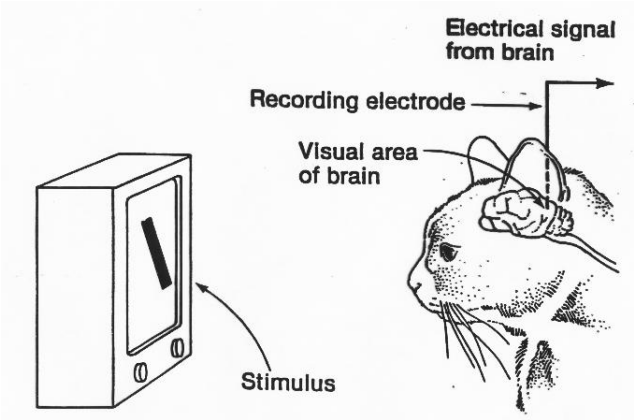


Figure: Famous Hubel and Wiesel cat experiment. (adopted from Hubel and Wiesel 1959).

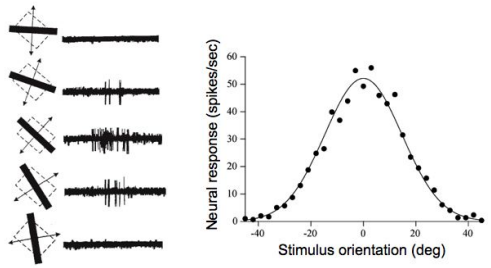


Figure: Neural response of simple cells. (adopted from Hubel and Wiesel 1968).

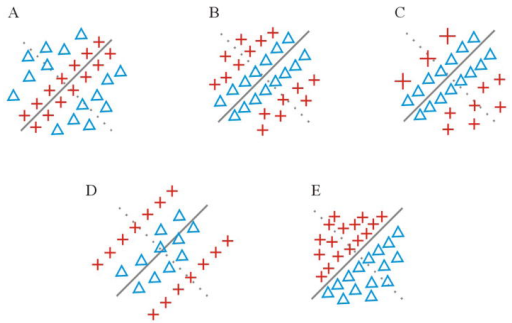


Figure: Simple receptive fields. (adopted from Hubel and Wiesel 1962).

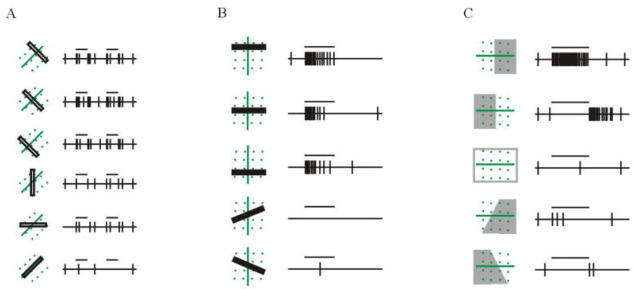


Figure: Three different types of complex receptive fields. (adopted from Hubel and Wiesel 1962).

Neurobiological foundations of CNNs

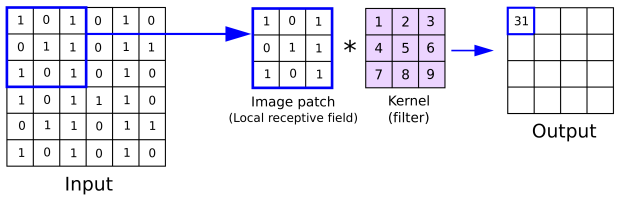


Figure: Example of a 2-dimensional matrix convolution.

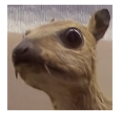
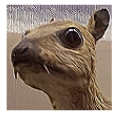
Original	Gaussian Blur	Sharpen	Edge Detection
$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$
			

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

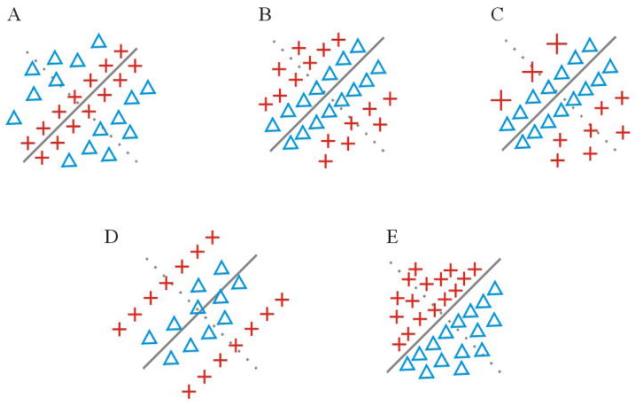


Figure: Simple receptive fields. (adopted from Hubel and Wiesel 1962).

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

Neurobiological foundations of CNNs

The computations that the **complex cells** perform are similiar to the **max pooling** operation.

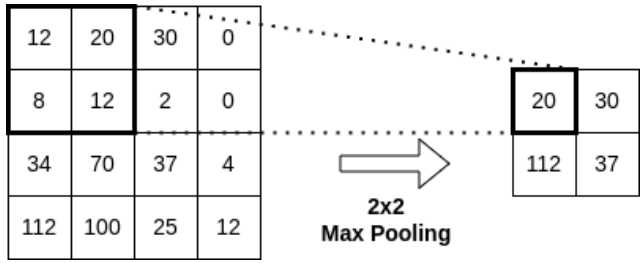


Figure: Example of a 2x2 max pooling matrix operation.

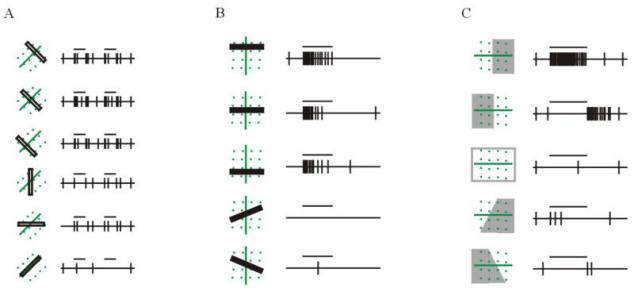


Figure: Three different types of complex receptive fields. (adopted from Hubel and Wiesel 1962).

Simple and complex cells in the auditory cortex

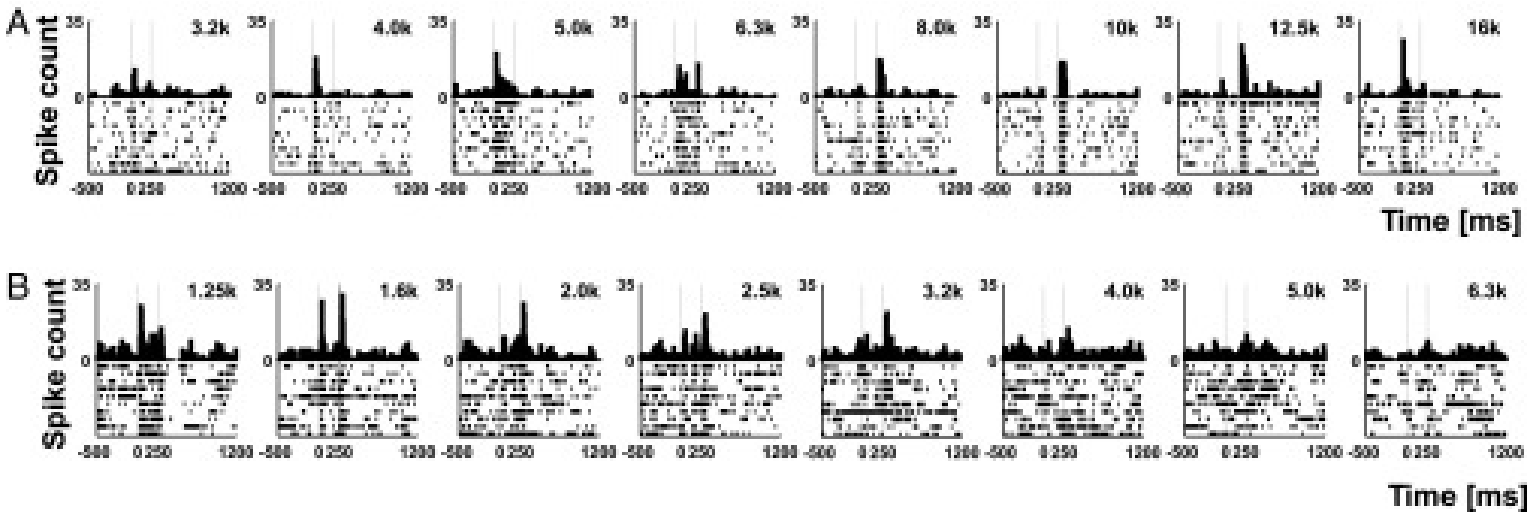


Figure: Responses of single neurons in primary auditory cortex (A1) of rhesus monkeys to band-passed noise (BPN) bursts centered at particular frequencies.. (Adopted from Tian et al. 2013)

Auditory cortex also contains **simple (and complex) cells** with two-dimensional receptive fields that are similar to those in the visual cortex but which operate in the **spectrotemporal domain** instead.

Explanatory properties of CNNs

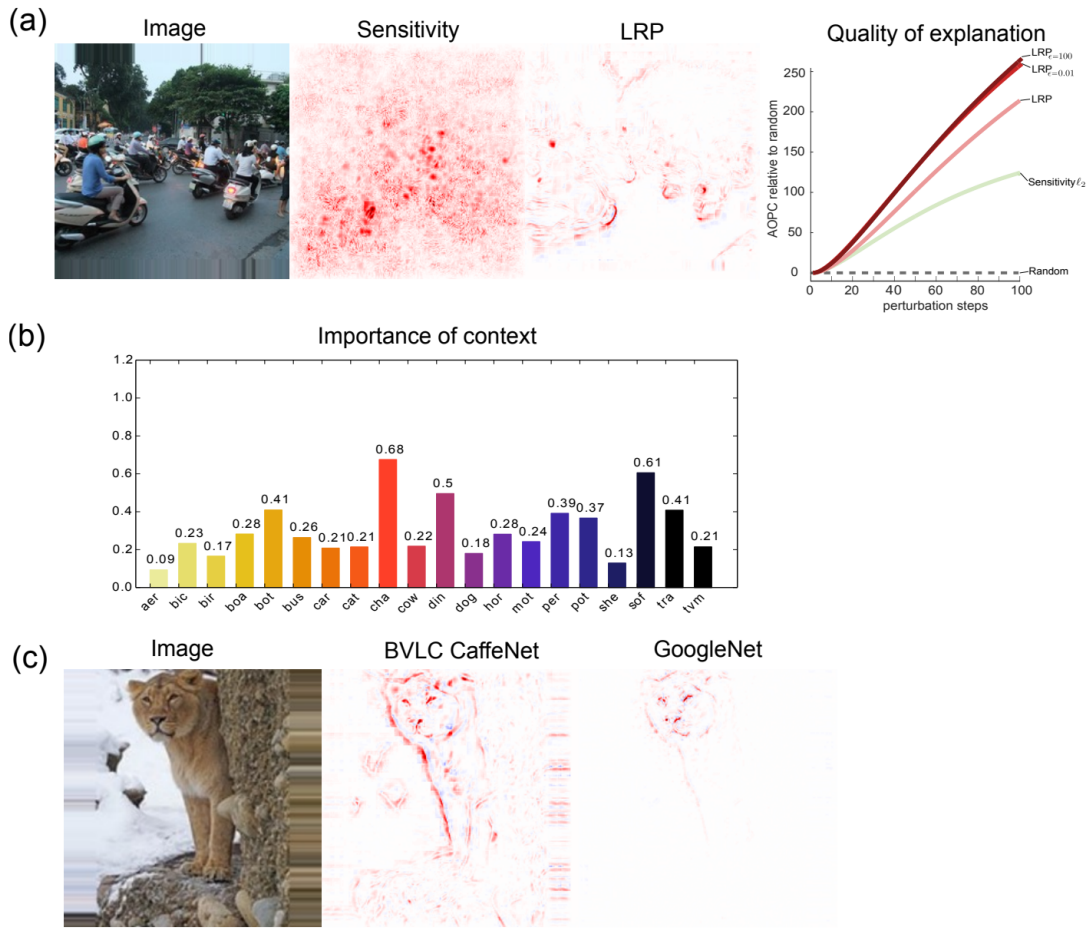


Figure: Results of sensitivity-based and relevance-based explainability methods. (Based on Samek 2016)

Dataset

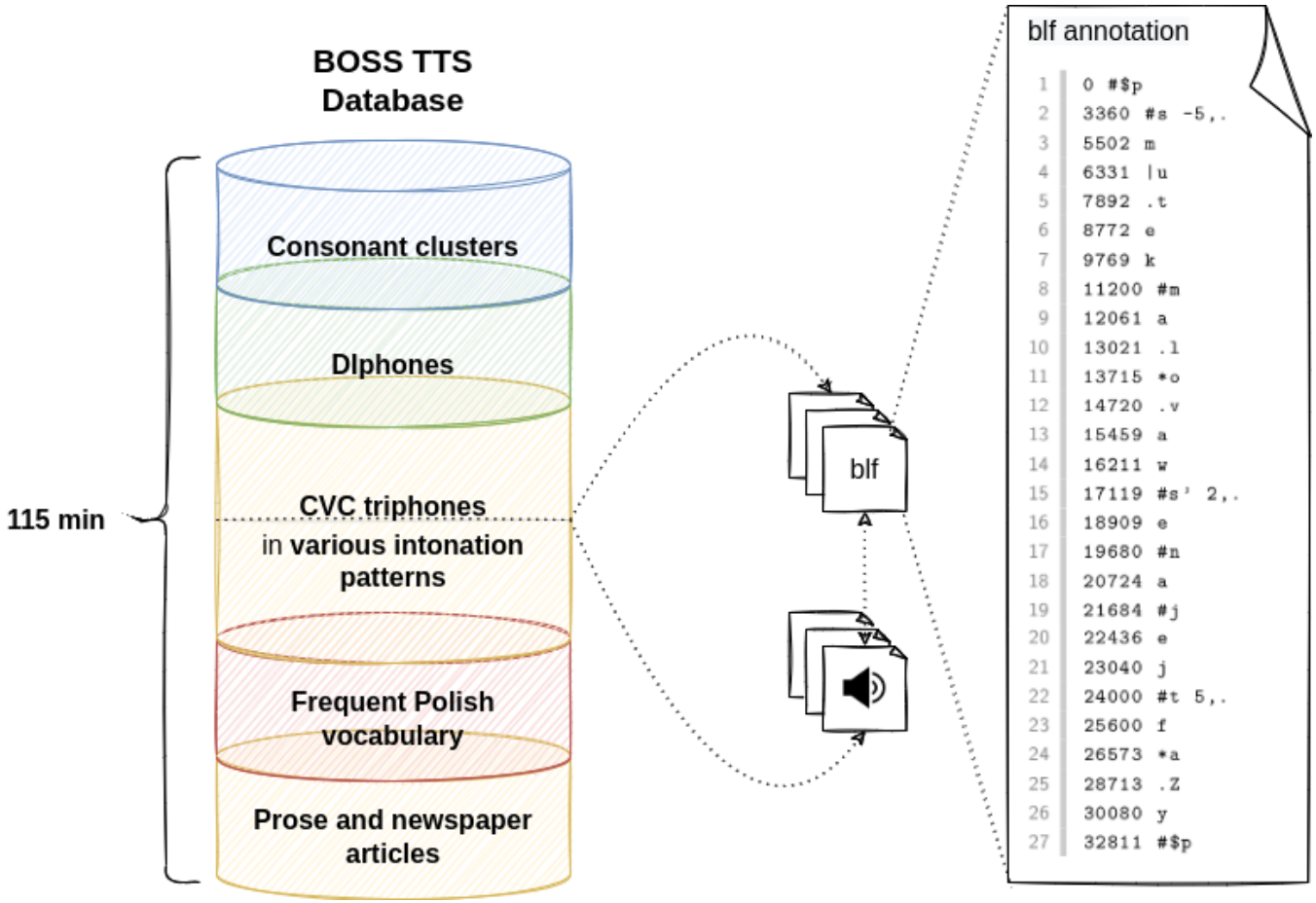
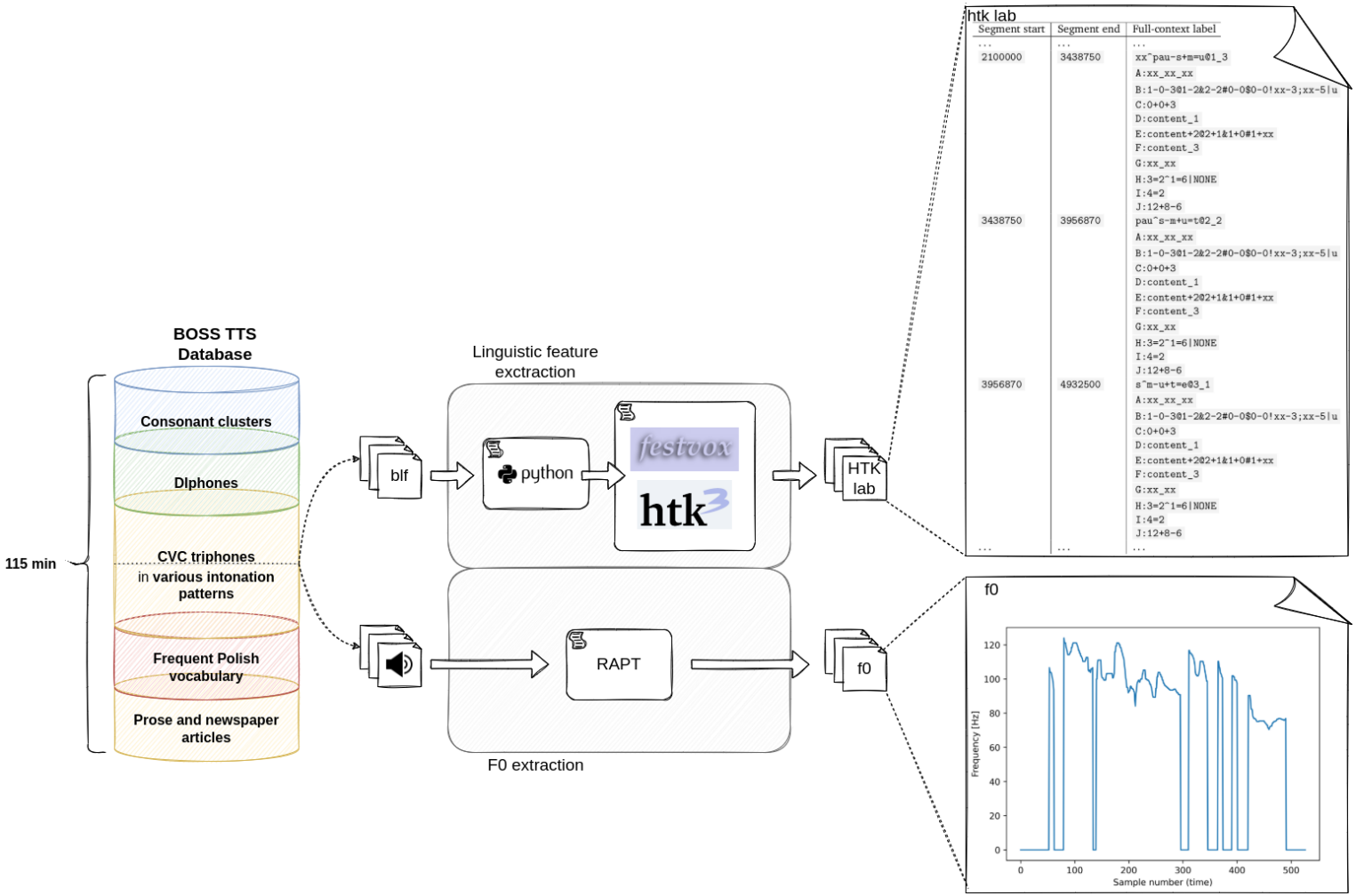
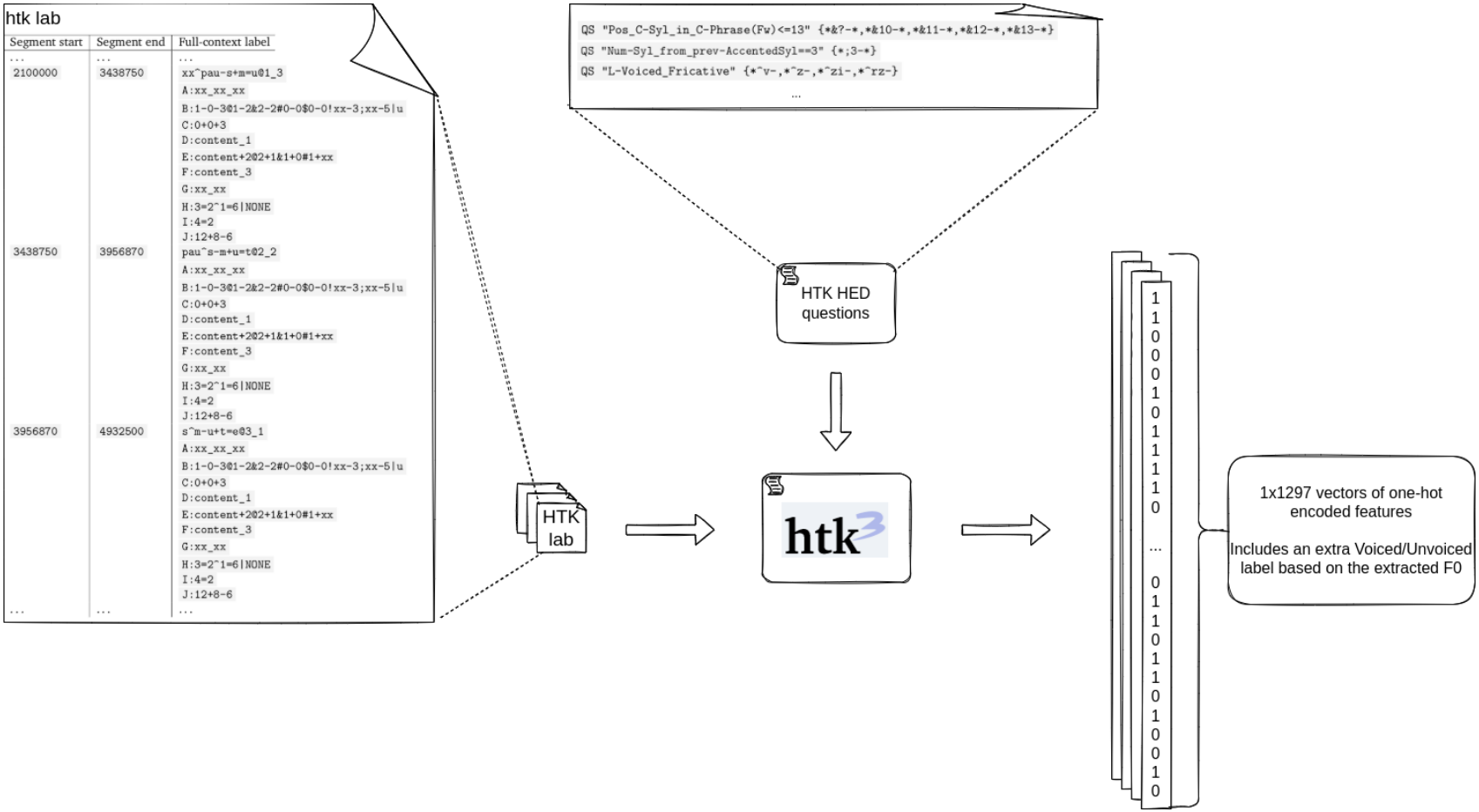


Figure: Speech corpus built originally for the purpose of the Polish BOSS unit selection synthesizer (Demenko, Bachan, Möbius 2008; Demenko, Klessa, Szymański, Breuer, Hess 2010).

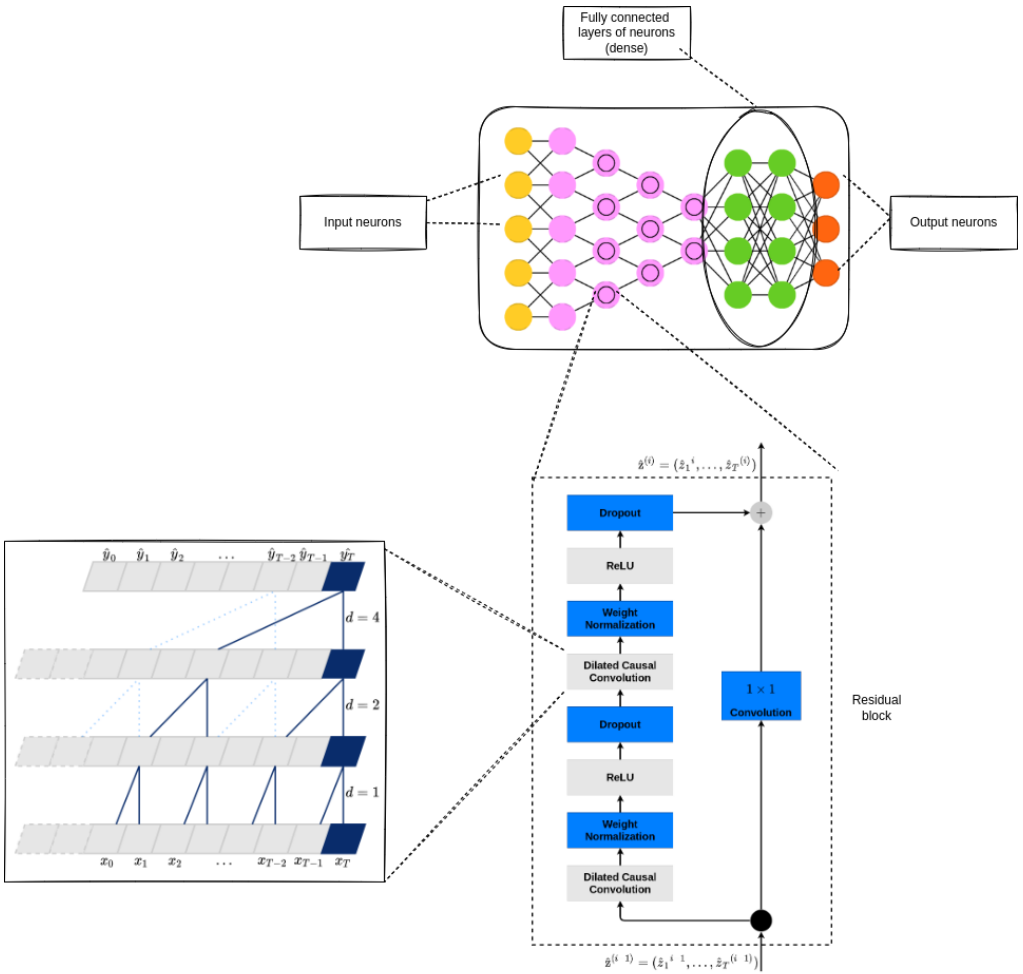
Data preprocessing



Feature extraction



Model implementation



TCN parameters

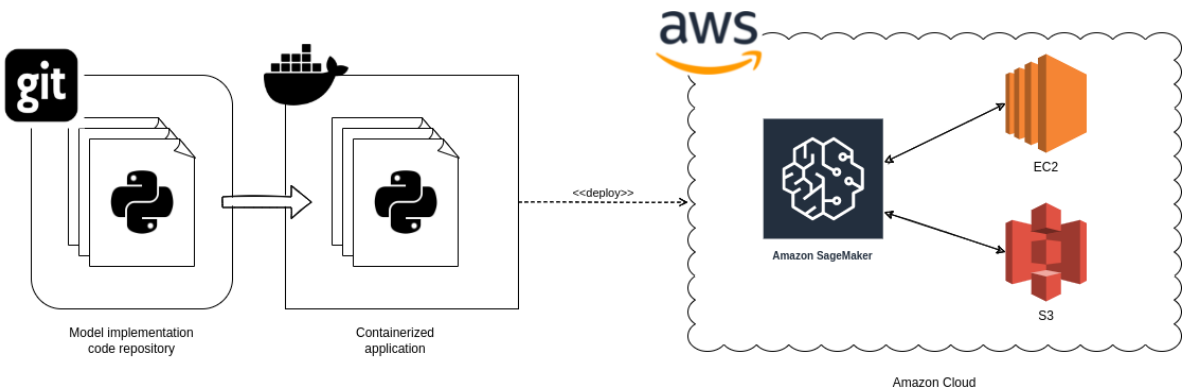
- 6 residual layers,
- 64 convolution filters of length 2
- 2^n dilations ([1, 2, 4, 8, 16, 32])
- 2 final fully connected dense layers (ReLU activation; sizes 64 and 1)

Complete code repository

✓ `https://github.com/mrslacklines/intonation_synthesis`

`github.com/philipperemy/keras-tcn`

Model training infrastructure



- 64 CPUs; 488GB
- 8 NVIDIA Tesla V100 GPUs; 128GB

Memory requirements

A single batch of data, which is a $64 \times 1900 \times 1297$ vector of 8-byte boolean values, occupies 1.2617216 gigabytes of memory, and the model comprises of a total of 449,409 parameters (446,337 trainable and 3,072 non-trainable).

TensorFlow
Keras
python
docker
aws

Training parameters:

- ADAM optimizer
- initial learning rate of 0.1,
- $\beta_1 = 0.9$,
- $\beta_2 = 0.999$,
- $\varepsilon = 1e - 07$.
- **Loss metric**: Mean Squared Error (MSE)
- 200 epochs
- random 8:1:1 dataset split

F0 inference and feature relevance analysis

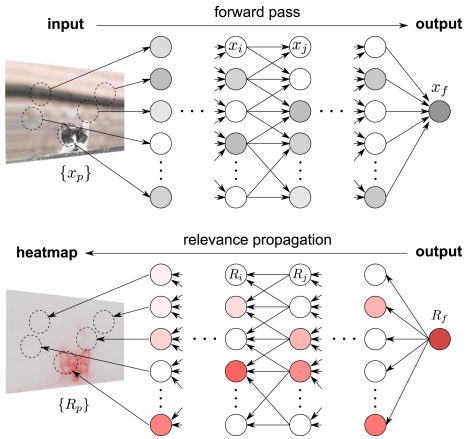
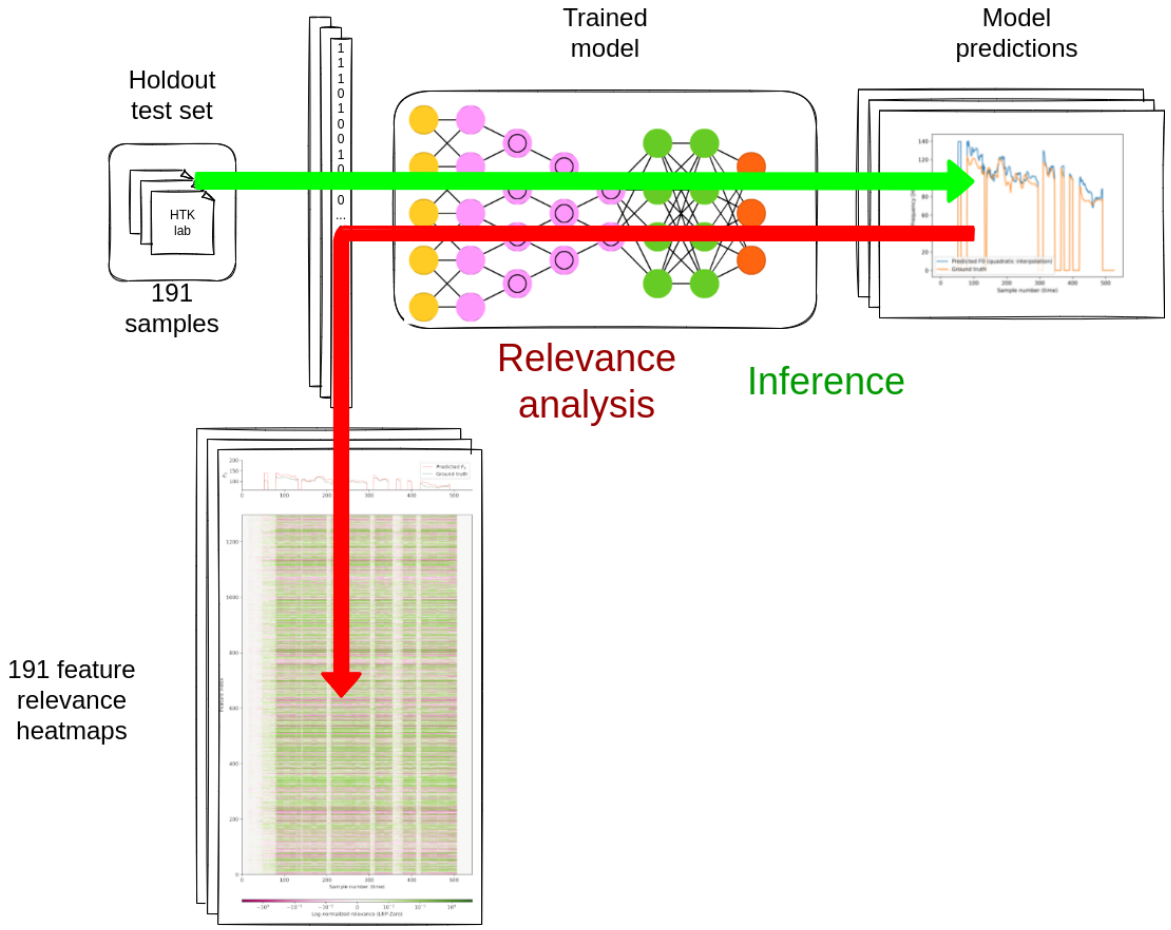
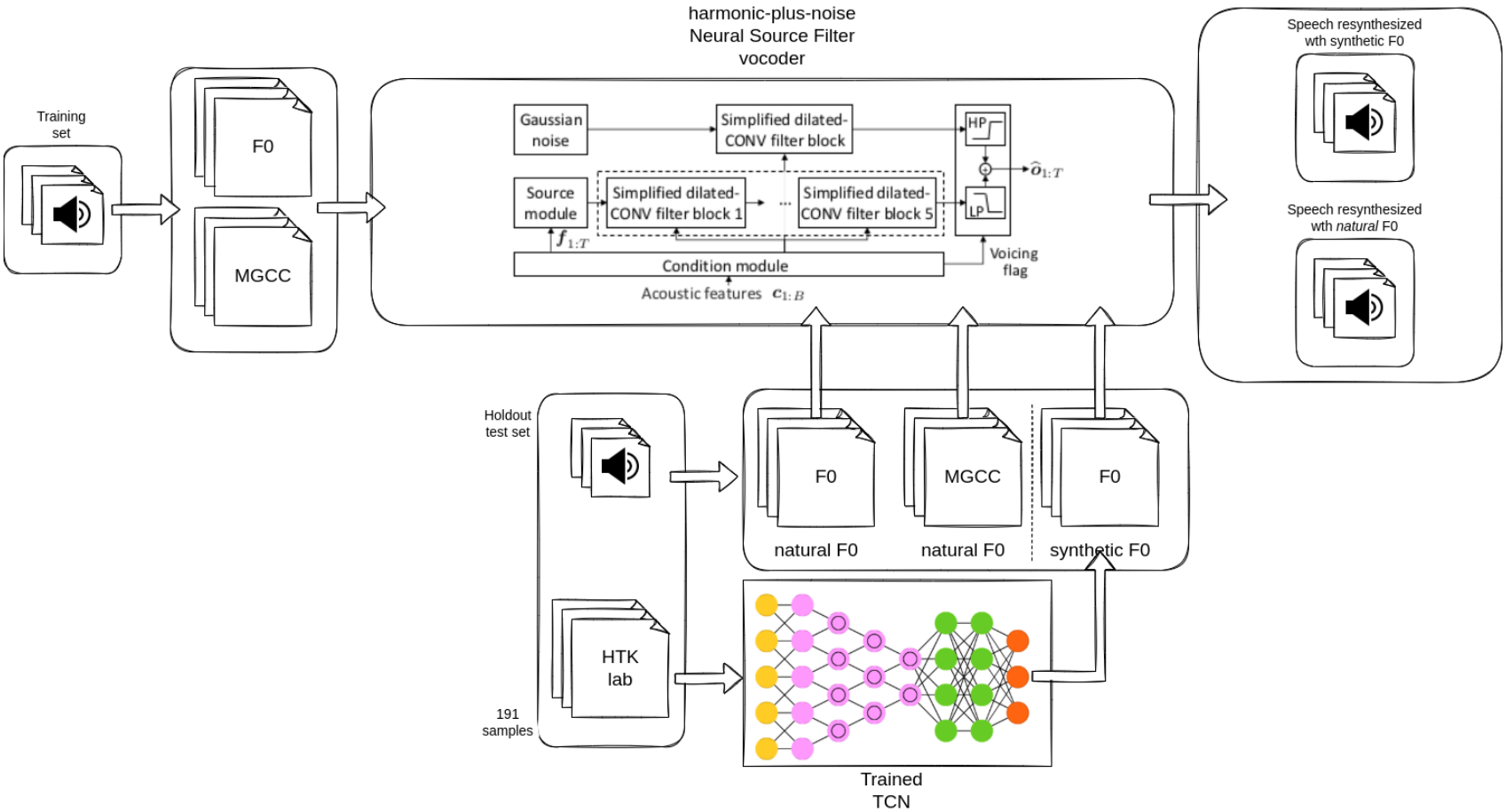


Figure: Computational flow of deep Taylor decomposition. (Adopted from Montavon 2017).

INNvestigate Neural Networks! (Alber et al. 2019)
`github.com/albermax/innvestigate`

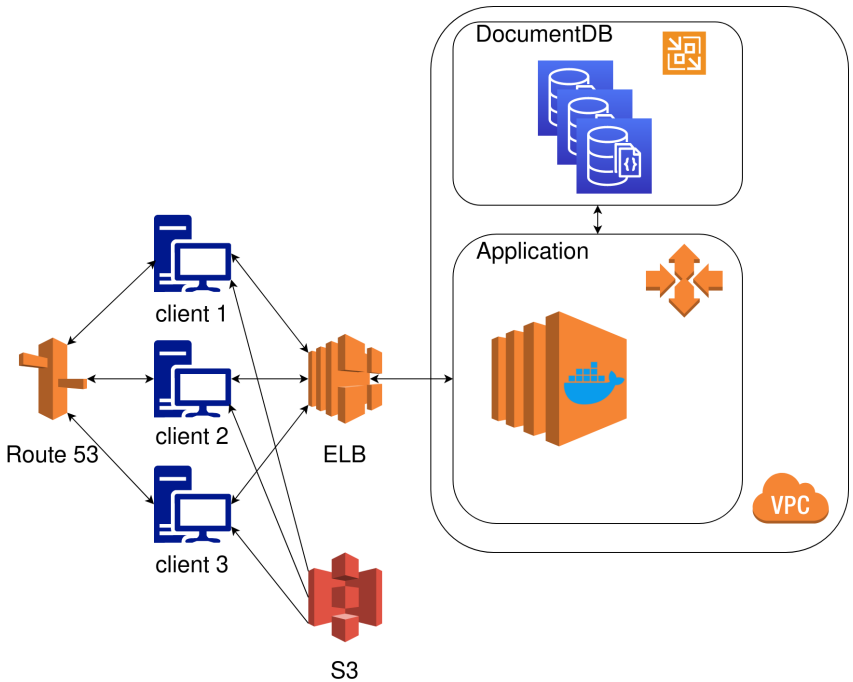


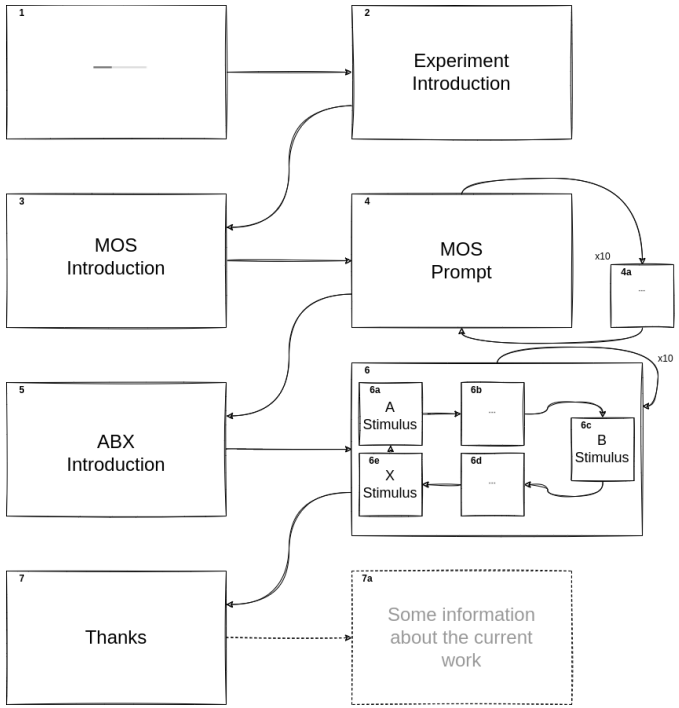
Resynthesis



Neural Source Filter Vocoder from (X. Wang et al. 2019)

Perceptual evaluation experiment





jspsych

Flask
web development, one drop at a time

Available at:
`fonetyka.cudaniewidy.org/experiment`
Code at:
`/github.com/mrslacklines/listening_experiments`

F0 inference results

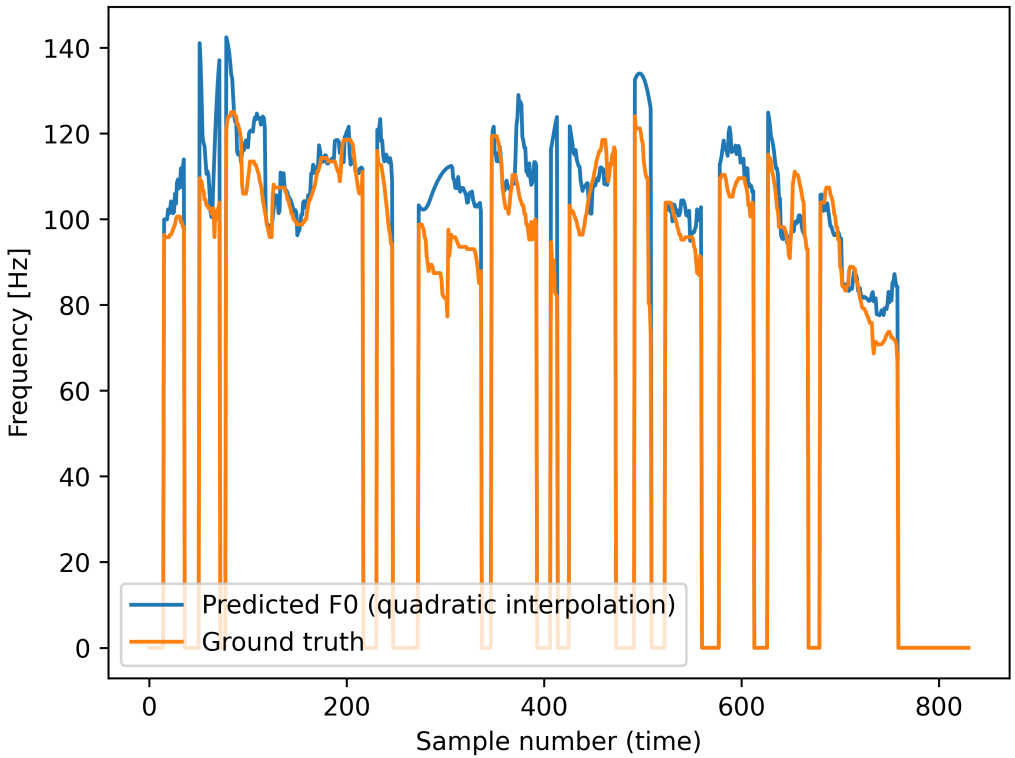


Figure: Result of F_0 prediction for "Lokatorzy znaleźli się w podbramkowej sytuacji i musieli się wyprowadzić" (*The tenants found themselves in a difficult situation and had to move out*).

F0 inference results

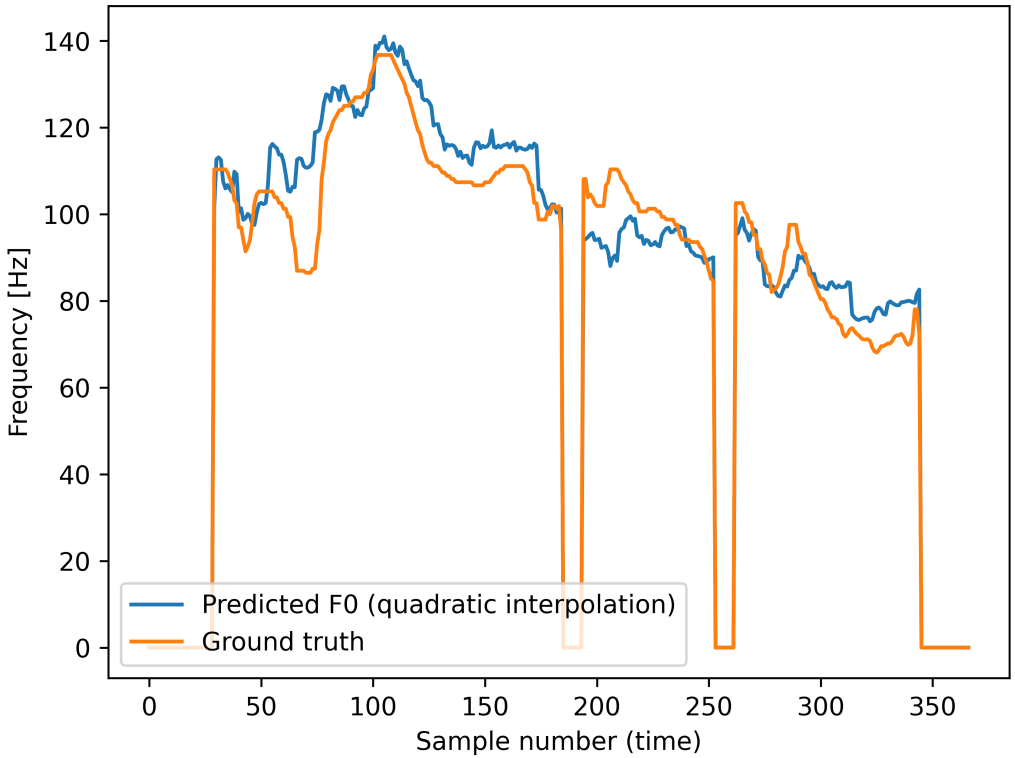


Figure: Result of F_0 prediction for "Powodzenie nie jest gwarantowane" (*Success is not guaranteed*).

F0 inference results

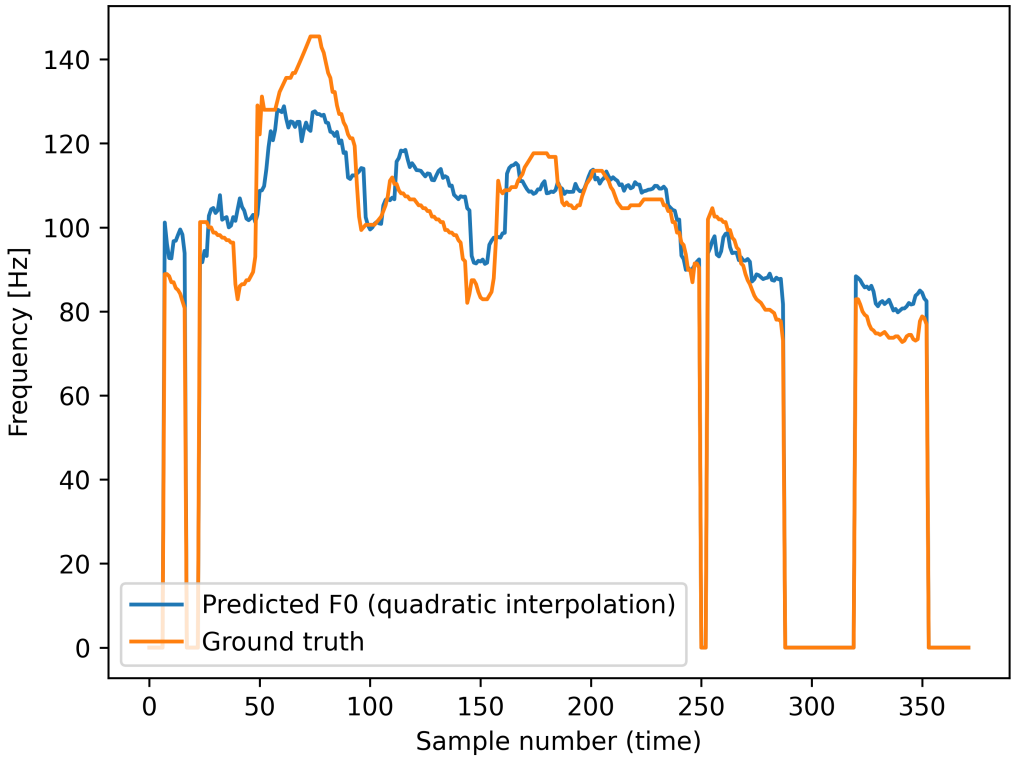


Figure: Result of F_0 prediction for "Gaduła była bardzo nieznośna" (*Gabby was very annoying.*).

F0 inference results

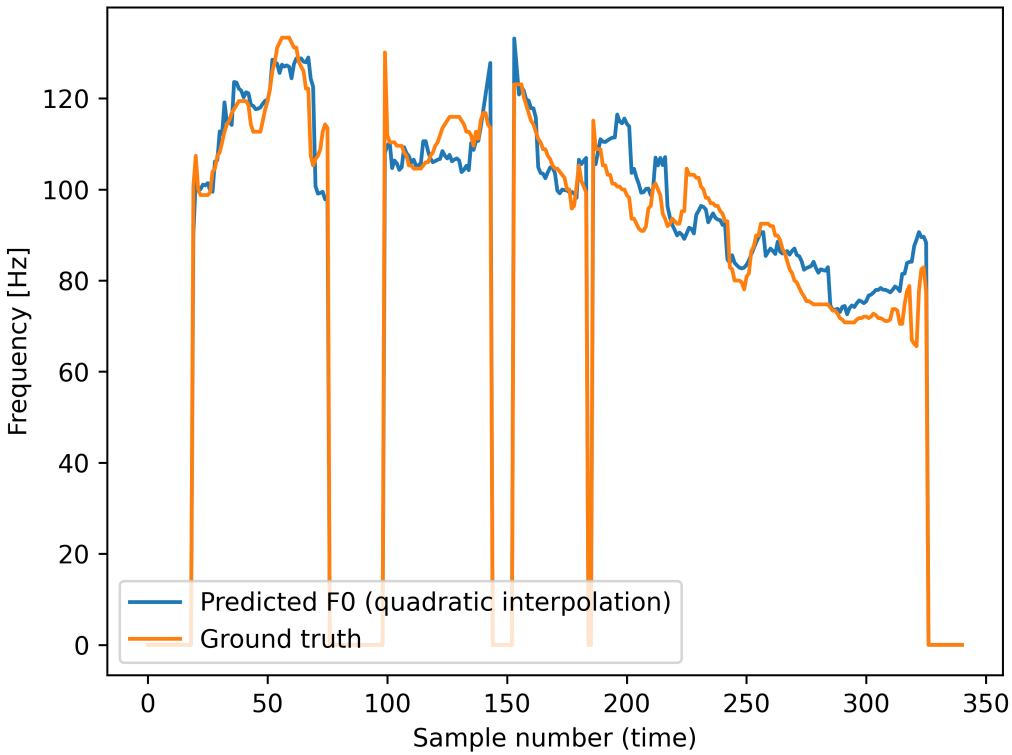


Figure: Result of F_0 prediction for "Może przyniosą też gorzałę" (*Maybe they will bring booze too.*).

F0 inference results

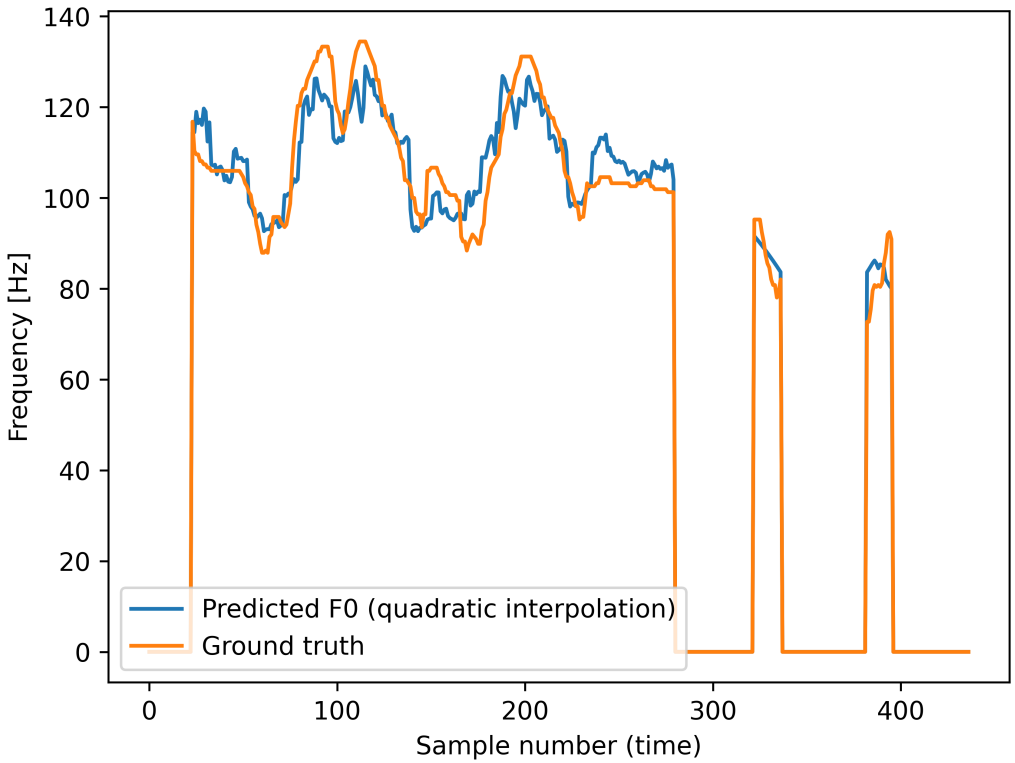


Figure: Result of F_0 prediction for "To jest ważna godzina dla nas wszystkich" (*This is an important hour for all of us*).

F0 inference results

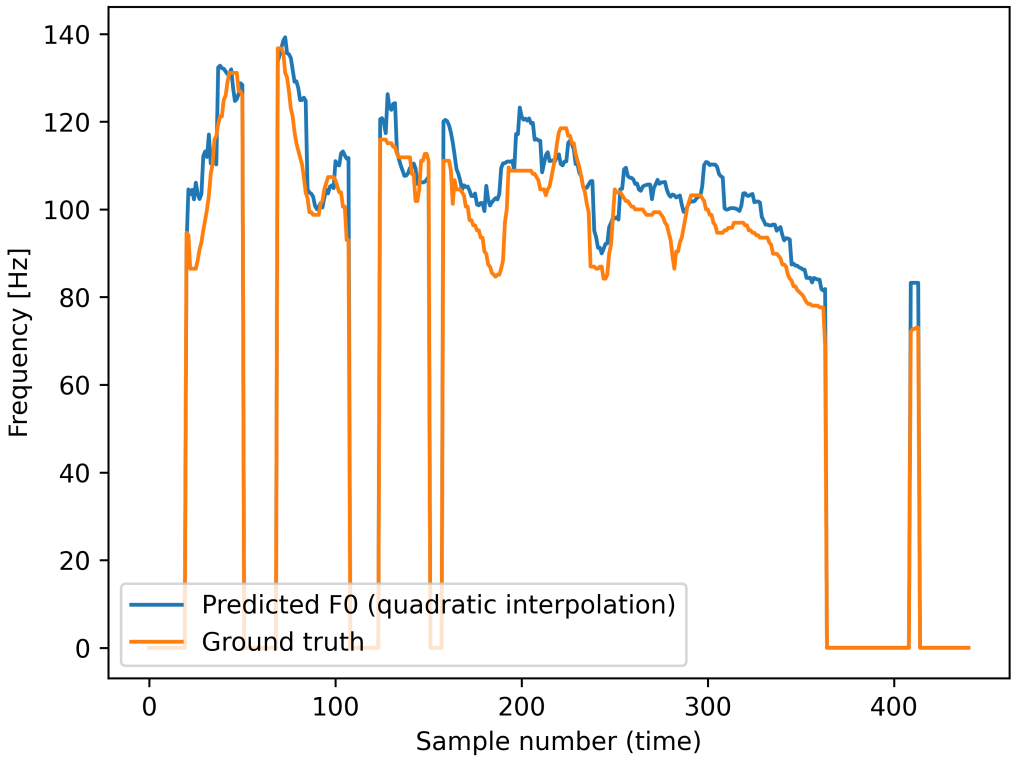


Figure: Result of F_0 prediction for "Myślę, że chleb razowy będzie najlepszy" (*I think that a wholemeal bread will be the best*).

F0 inference results

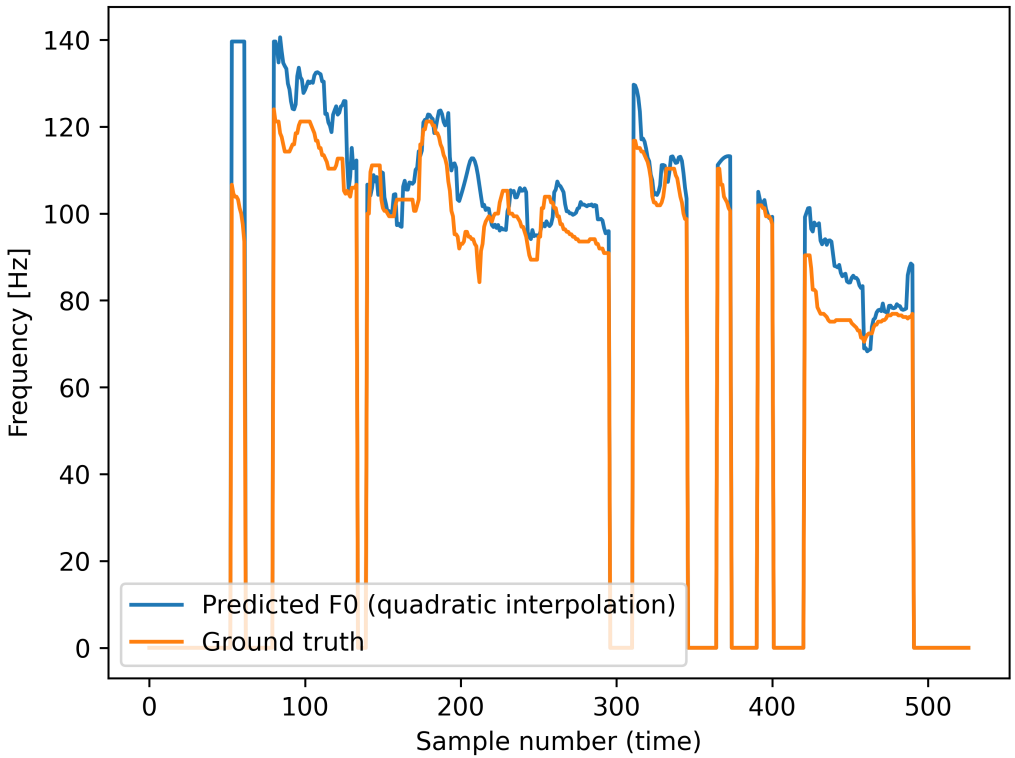


Figure: Result of F_0 prediction for "Słyszałam odgłos zbliżającego się pociągu" (*I heard the sound of an approaching train*).

F0 inference results

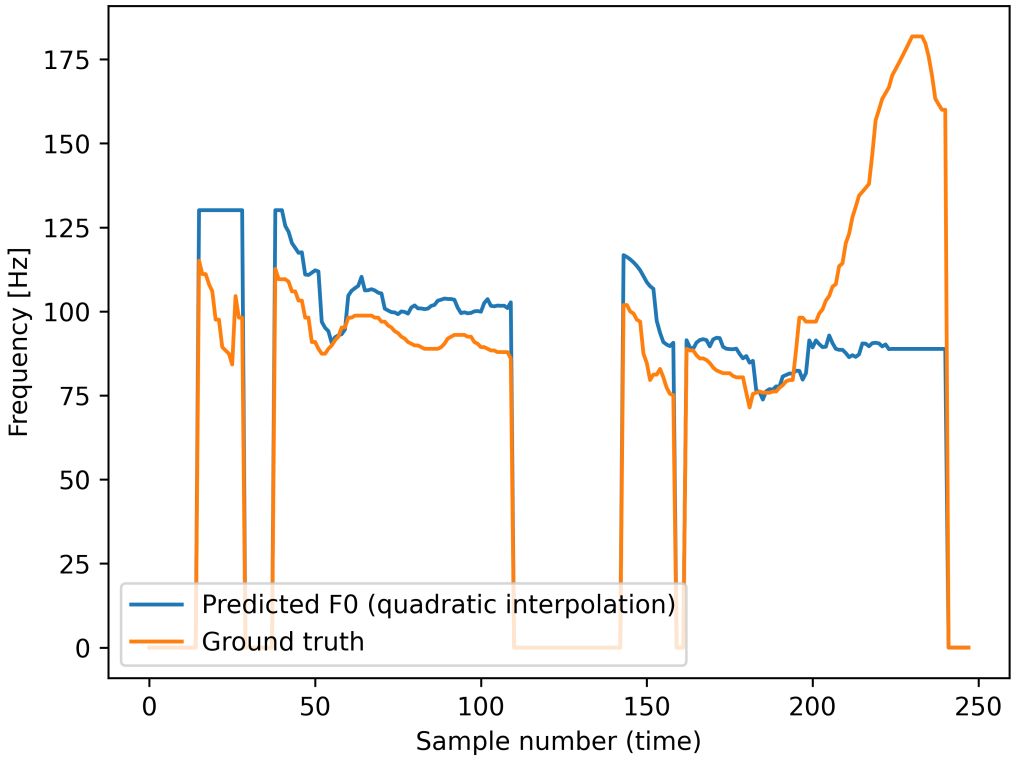


Figure: Result of F_0 prediction for "Czy to był łatwy dobór?" (*Was it an easy choice?*).

F0 inference results

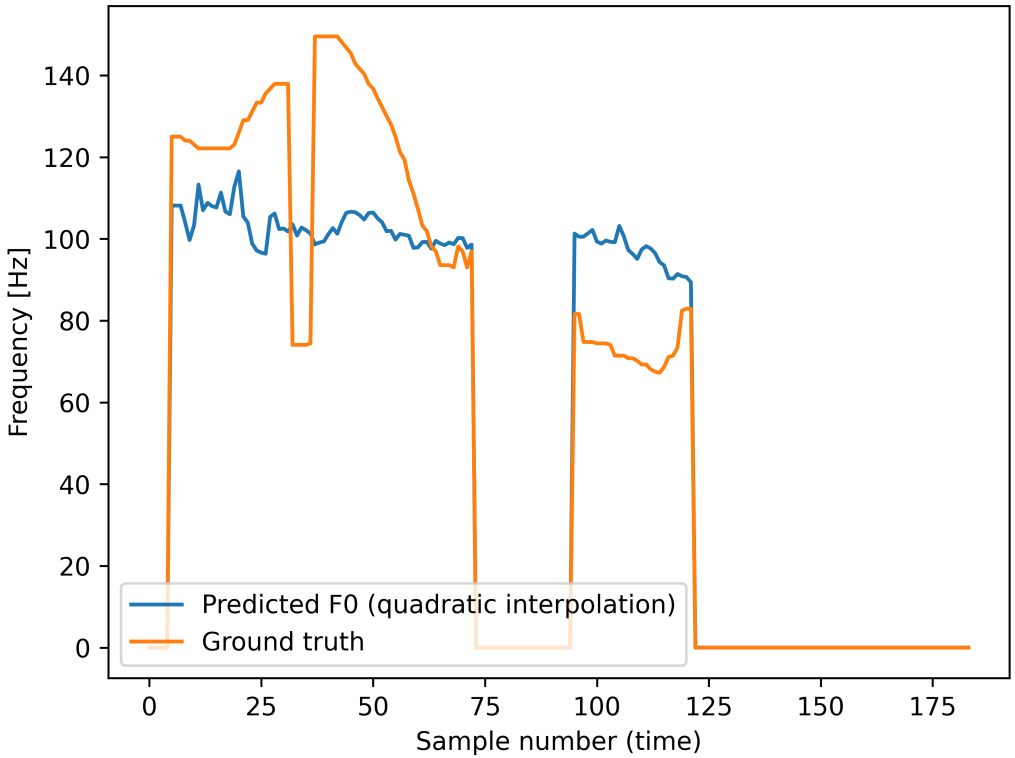


Figure: Result of F_0 prediction for "To Majka" (*This is Majka*).

F0 inference results

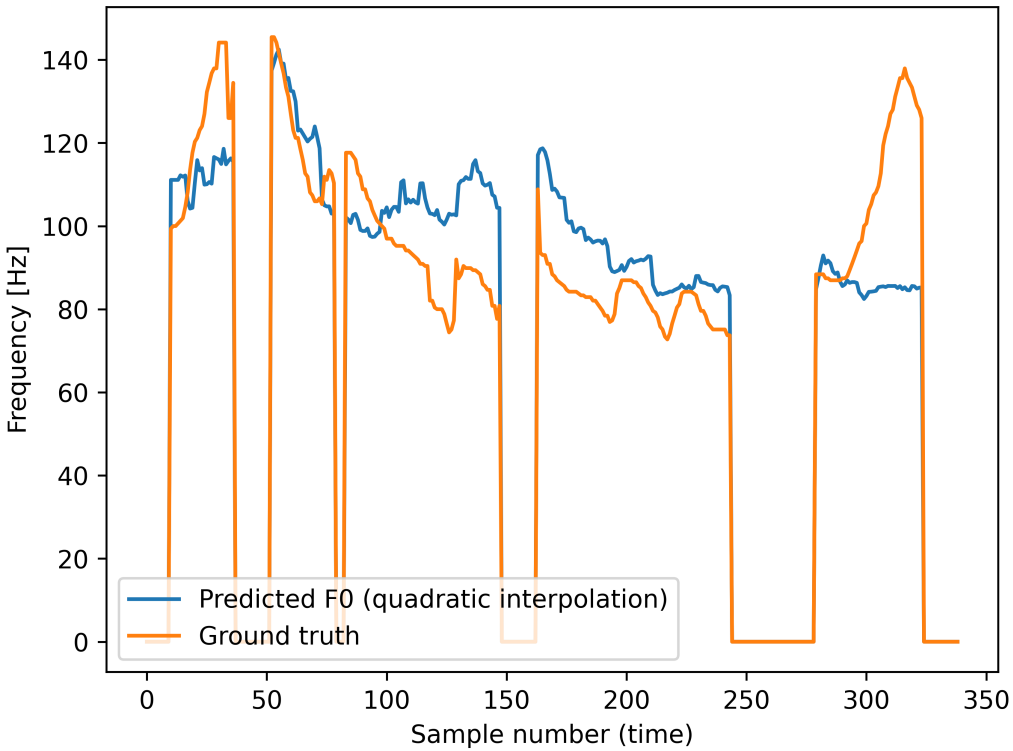


Figure: Result of F_0 prediction for "Na czym polega kandyzacja?" (*How does candying work?*).

Subjective evaluation results - MOS

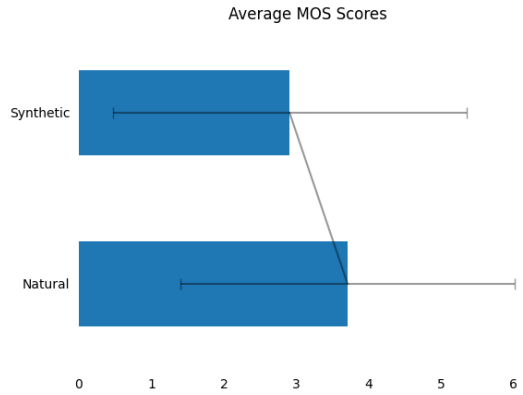


Figure: Mean Opinion Score-based evaluation results.

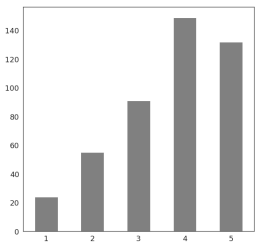


Figure: Mean Opinion Score-based evaluation total numbers of specific scores for natural stimuli.

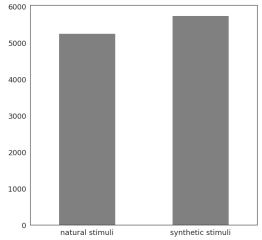


Figure: Mean Opinion Score-based evaluation mean response times for synthetic and natural stimuli.

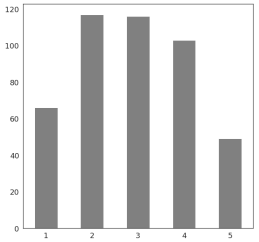


Figure: Mean Opinion Score-based evaluation total numbers of specific scores for synthetic stimuli.

Subjective evaluation results - MOS

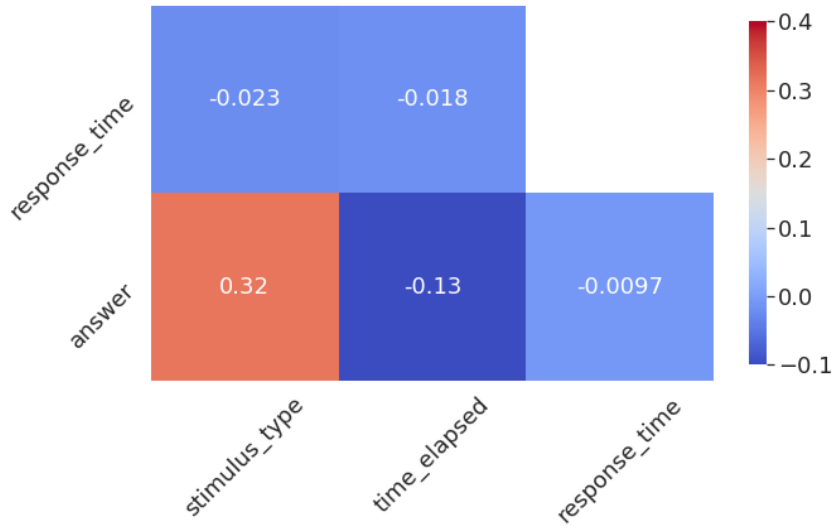
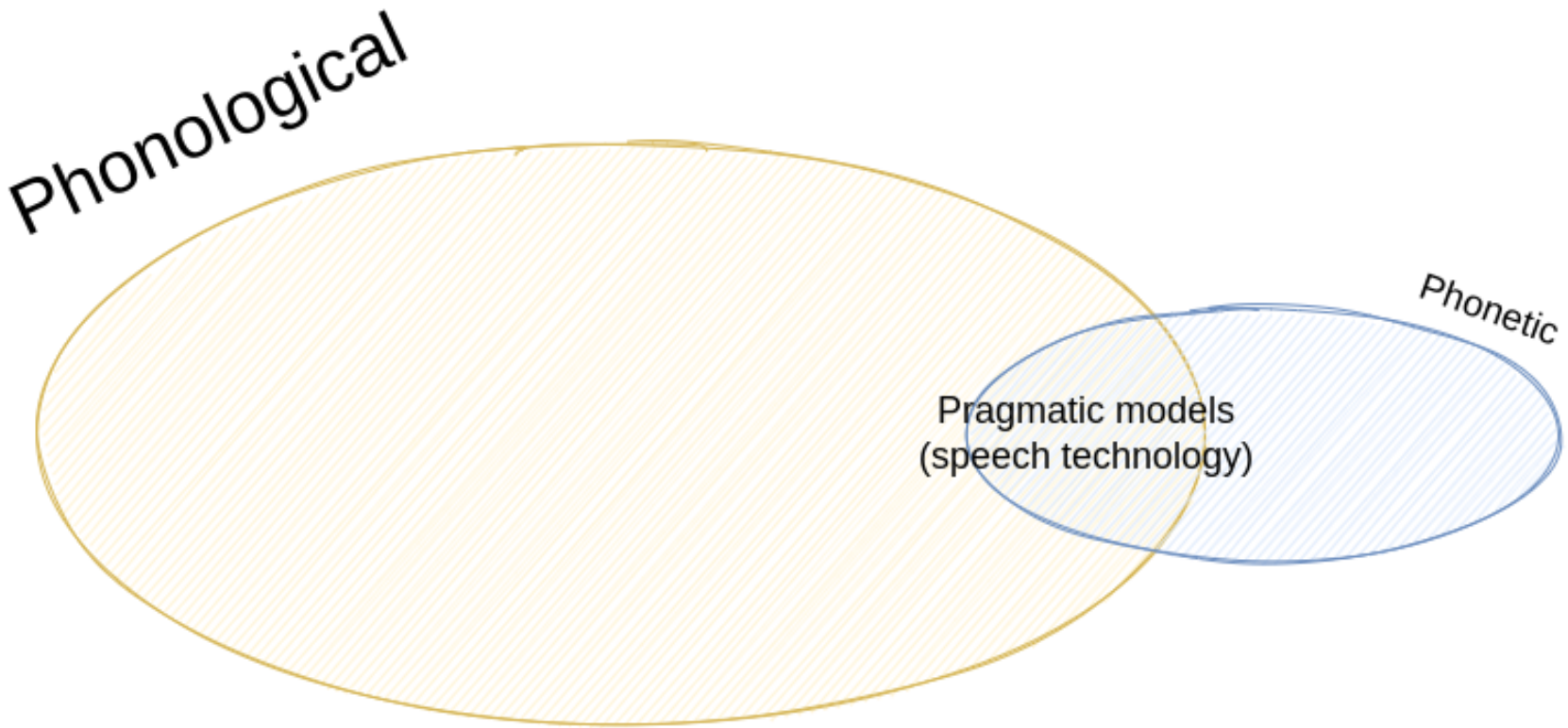


Figure: Mean Opinion Score-based evaluation parameters correlation matrix.

Thank you.

Intonation models



Speech synthesis - Wavenet

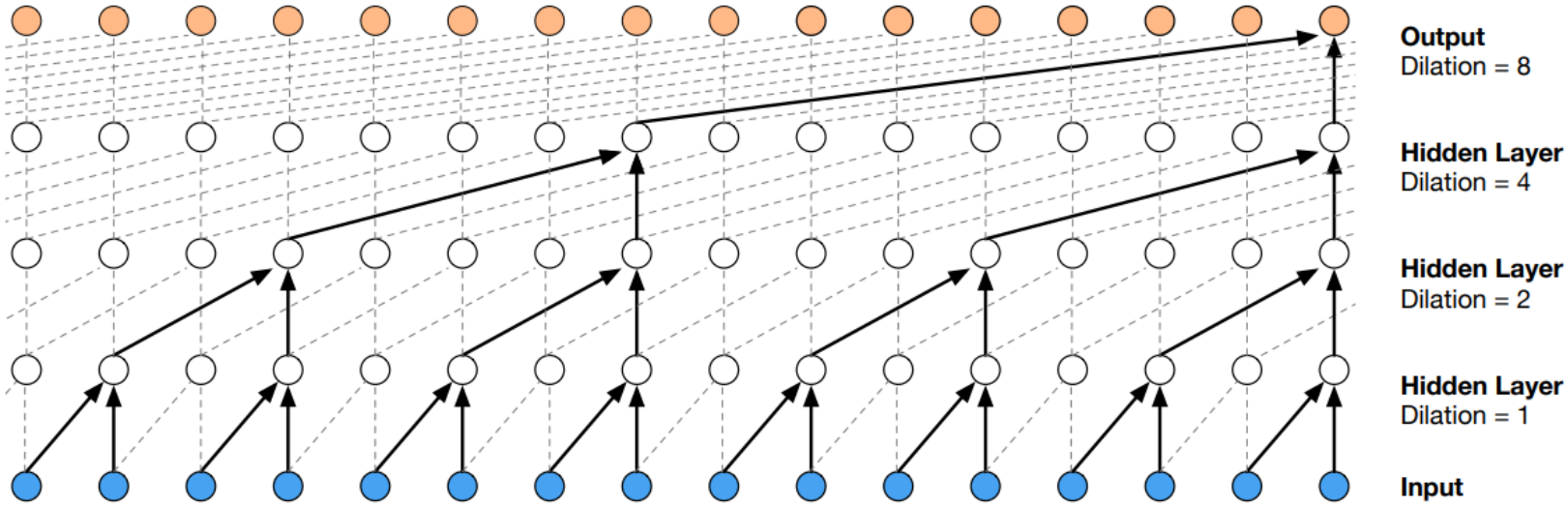


Figure: Dilated causal convolutions. (Adopted from the original WaveNet paper).

The causality is expressed through the joint probability of the modeled waveform $\vec{x} = \{x_1, \ldots, x_T\}$ being factorized as a product of conditional probabilities of all previous timesteps (van den Oord 2016), i.e.:

$$p(\vec{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) \quad (1)$$

Speech synthesis - Wavenet

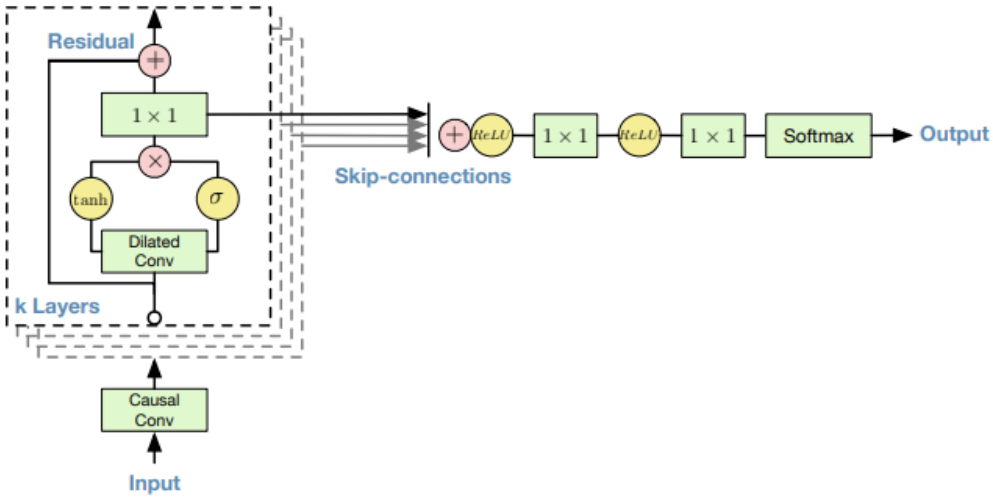


Figure: Residual and skip connections from a stack of k gated convolutional layers (Adopted from the original WaveNet paper).

Gated convolutional layers:

$$\vec{z} = \tanh\left(W_{f,k} * \vec{x}\right) \odot \sigma\left(W_{g,k} * \vec{x}\right), \quad (2)$$

where $*$ denotes a convolution operator, $\odot$ denotes an element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, k is the layer index, f and g denote filter and gate, respectively, and W is a learnable convolution filter.

Dataset

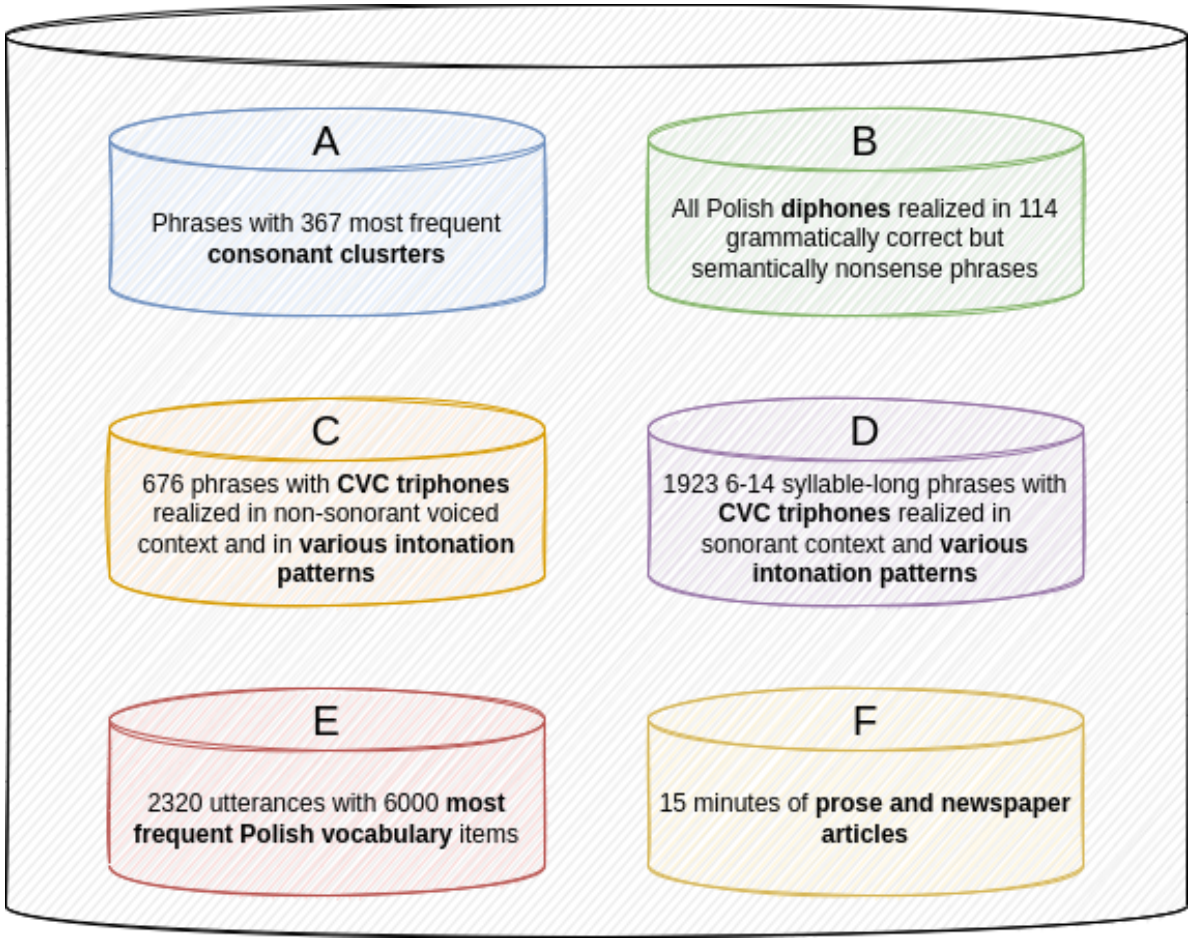


Figure: Speech corpus built originally for the purpose of the Polish BOSS unit selection synthesizer (Demenko, Bachan, Möbius 2008; Demenko, Klessa, Szymański, Breuer, Hess 2010).

Stress and accent type labels

%	rising accent realized by F_0 rise on post accented syllable/syllables or F_0 interval between accented and post accented vowels
’	rising accent realized by F_0 change (rise on accented syllable)
"	falling accent realized by F_0 fall on post accented syllable/syllables or F_0 interval between accented and post accented vowels
&	falling accent realized by F_0 change (fall on accented syllable)
|	rising-falling accents with rise-fall shape of F_0 movement on accented vowel
*	level accent realized by F_0 interval between preaccented and accented vowels; near zero slope of fundamental frequency
<	level accent realized only by differences in duration between preaccented, accented and postraccented vowels

Figure: Stress and accent labels used in the original Polish BOSS speech corpus.

Stress and accent type labels

`-5,.`	Intonation on the first word in a sentence with falling accent F (or level accent L). In most cases it is used for declarative sentences or wh-questions. Mark on the first phoneme of the first word in the sentence
`-5,?`	Intonation on the first word in a sentence with rising accent R. It can be used in different complex sentences. Mark on the first phoneme of the first word in the sentence
`5,.`	Intonation on the last word in sentence with falling accent F (or level accent L). In most cases it is used for declarative sentences or wh-questions. Mark on the first phoneme of the last word in the sentence
`5,?`	Intonation on the last word in a sentence with rising accent R. In most cases it is used for yes-no questions. Mark on the first phoneme of the last word in the sentence
`5.!`	Intonation on the last word in a sentence with falling accent F. In most cases it is used for exclamatory sentences. Mark on the first phoneme of the last word in the sentence.
`2,?`	Intonation on the last word in the phrase with rising accent R. In most cases it is used for continuation phrases. Mark on the first phoneme of the last word in the phrase.
`2,.`	Intonation on the last word in the phrase with falling accent F (or level accent L). In most cases it is used in declarative phrases in complex sentences. Mark on the first phoneme of the last word in the sentence.

Figure: Prosodic phrase boundary labels used in the original Polish BOSS speech corpus.

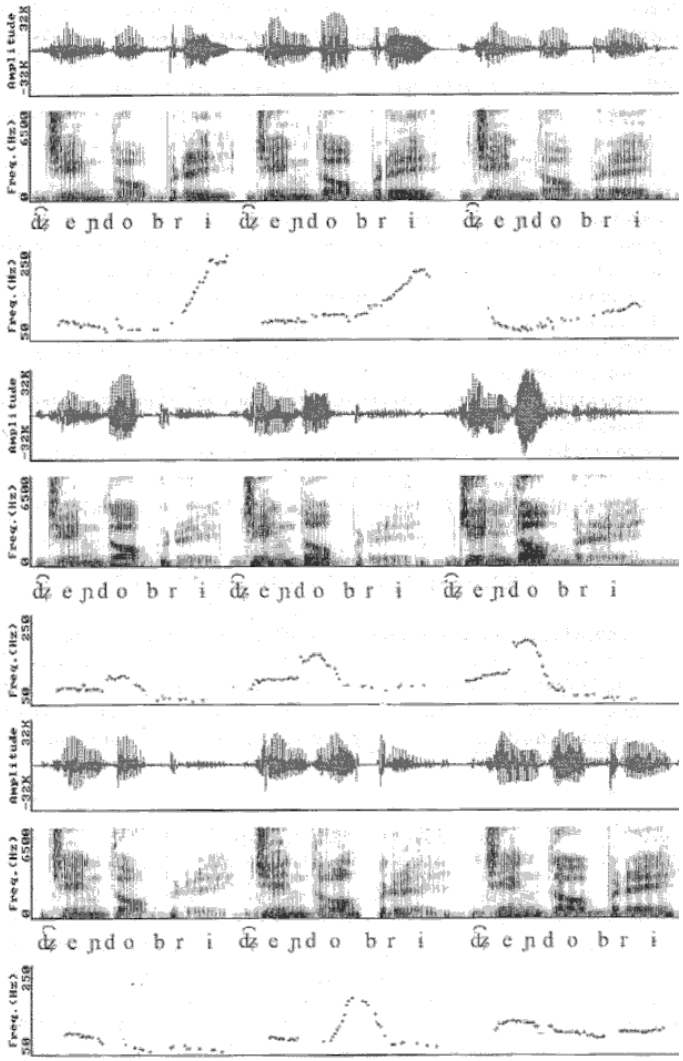
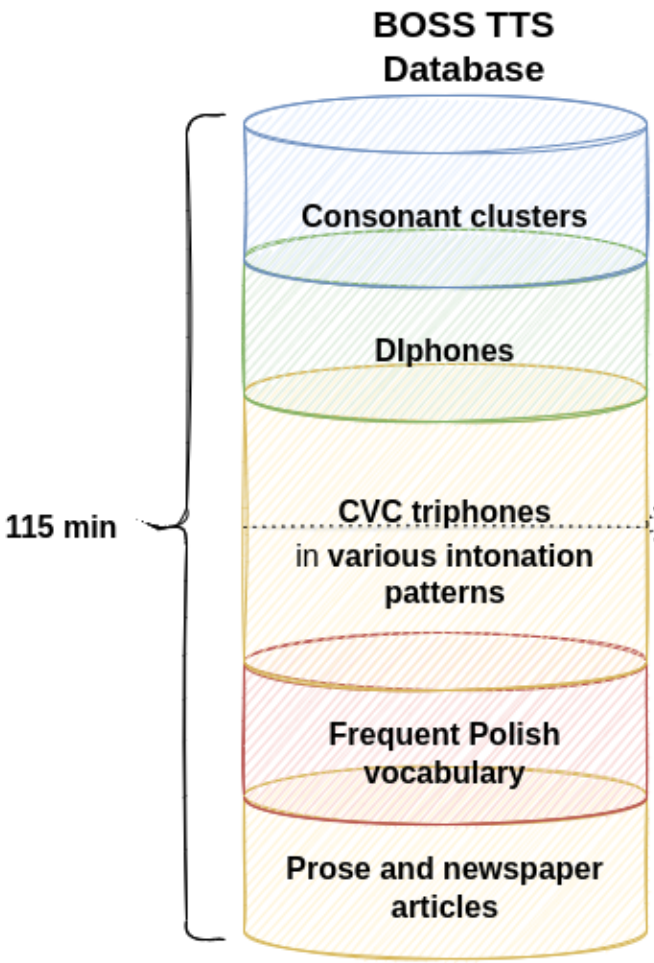


Figure: Acoustic realizations of the 9 different accents. Adopted from (Demenko 1999).

Dataset



Stress and accent type labels

-5,.	Intonation on the first word in a sentence with falling accent F (or level accent L). In most cases it is used for declarative sentences or wh-questions. Mark on the first phoneme of the first word in the sentence
-5,?	Intonation on the first word in a sentence with rising accent R. It can be used in different complex sentences. Mark on the first phoneme of the first word in the sentence
5,.	Intonation on the last word in sentence with falling accent F (or level accent L). In most cases it is used for declarative sentences or wh-questions. Mark on the first phoneme of the last word in the sentence
5,?	Intonation on the last word in a sentence with rising accent R. In most cases it is used for yes-no questions. Mark on the first phoneme of the last word in the sentence
5.!	Intonation on the last word in a sentence with falling accent F. In most cases it is used for exclamatory sentences. Mark on the first phoneme of the last word in the sentence.
2,?	Intonation on the last word in the phrase with rising accent R. In most cases it is used for continuation phrases. Mark on the first phoneme of the last word in the phrase.
2,.	Intonation on the last word in the phrase with falling accent F (or level accent L). In most cases it is used in declarative phrases in complex sentences. Mark on the first phoneme of the last word in the sentence.

Stress and accent type labels

%	rising accent realized by F_0 rise on post accented syllable/syllables or F_0 interval between accented and post accented vowels
'	rising accent realized by F_0 change (rise on accented syllable)
"	falling accent realized by F_0 fall on post accented syllable/syllables or F_0 interval between accented and post accented vowels
&	falling accent realized by F_0 change (fall on accented syllable)
|	rising-falling accents with rise-fall shape of F_0 movement on accented vowel
*	level accent realized by F_0 interval between preaccented and accented vowels; near zero slope of fundamental frequency
<	level accent realized only by differences in duration between preaccented, accented and postraccented vowels

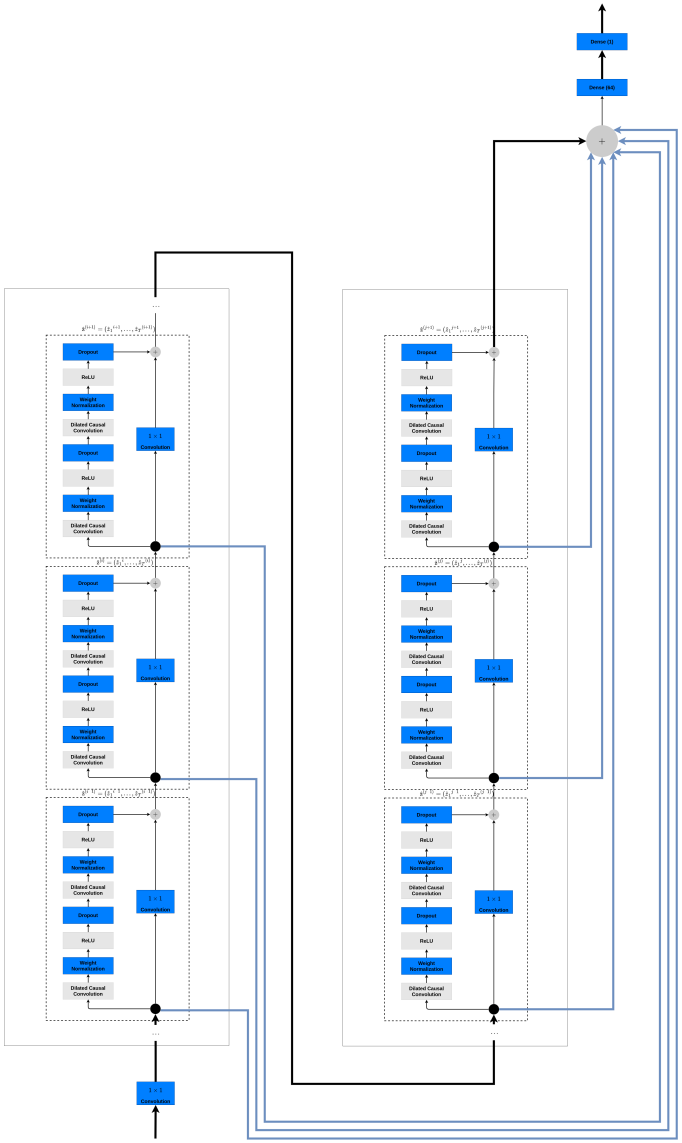
Feature set

Question type	Segments
Vowel	{i, y, e, a, o, u, schwa}
Consonant	{gs, p, b, t, d, k, g, ki, gi, f, v, s, si, z, zi, sz, rz, x, c, dz, cz, drz, ci, dzi, m, n, ni, ng, l, r, w, ww, j, jj}
Stop	{gs, p, b, t, d, k, g}
Nasal	{ww, jj, m, n, ni, ng}
Fricative	{f, v, s, si, z, zi, sz, rz, x}
Front	{e, i, y, f, v, p, b, m, w, ww}
Central	{schwa, a, t, d, s, si, z, zi, n, r, l, t, d, sz, rz, cz, drz, c, dz, ci, dzi}
Back	{o, u, k, g, ki, gi, ng, x, gs}
Front Vowel	{e, i, y}
Central Vowel	{a, schwa}
Back Vowel	{o, u}
High Vowel	{i, y, u}
Medium Vowel	{e, o}
Low Vowel	{a}
Rounded Vowel	{o, u}
Unrounded Vowel	{a, e, i, y}
XVowel (e.g. AVowel)	{i, y, e, a, o, u, schwa}
Unvoiced Consonant	{gs, p, t, k, ki, f, v, s, sz, x, c, cz, ci}
Voiced Consonant	{b, d, g, gi, v, z, zi, rz, dz, drz, dzi, m, n, ni, ng, l, r, w, ww, j, jj}
Front Consonant	{f, v, f, p, b, m, w, ww}
Central Consonant	{t, d, s, si, z, zi, n, r, l, t, d, sz, rz, cz, drz, c, dz, ci, dzi}
Back Consonant	{gs, k, g, ki, gi, ng, x}
Fortis Consonant	{gs, cz, f, k, p, s, sz, t, ci, c, ki}
Lenis Consonant	{drz, v, g, b, rz, z, d, dzi, dz, gi, zi}
Neigther F or L	{m, n, ni, ng, l, r, w, ww, j, jj}
Voiced Stop	{b, d, g}
Unvoiced Stop	{p, t, k, gs}
Front Stop	{b, p}
Central Stop	{d, t}
Back Stop	{g, k, gs}
Voiced Fricative	{v, z, zi, rz}
Unvoiced Fricative	{f, s, si, sz, x}
Front Fricative	{f, v}

Figure: Segmental features for a quintphone-wide context.

Number of preceding/succeeding segments in the previous/current/next syllable is equal to/less than or equal to 0-7
Previous/current/next syllable is stressed
Previous/current/next syllable is accented
Previous/current/next syllable has accent X
(where X is one of the ToBI accents described above)
Number of preceding/succeeding segments in the next syllable is equal to/less than equal to 0-7
Forward/backward position of the current syllable in current word is equal to/less than or equal to 0-7
Forward/backward position of the current syllable in current phrase is equal to/less than or equal to 0-20
Number of stressed syllables before/after the current syllable in current phrase is equal to/less than or equal to 0-12
Number of accented syllables before/after the current syllable in current phrase is equal to/less than or equal to 0-12
Number of accented syllables before/after the current syllable in current phrase is equal to/less than or equal to 0-7
Number of syllables from previous/next stressed syllable is equal to/less than or equal to 0-5
Number of syllables from previous/next accented syllable is equal to/less than or equal to 0-16
Current syllable nucleus is a non-vowel, vowel, front vowel, central vowel, back vowel, high vowel, medium vowel, low vowel, rounded vowel, unrounded vowel, [i], [e], [a], [o], [u], [y], [schwa]
Number of syllables in the previous/current/next word is equal to/less than or equal to 0-7
Forward/backward position of the current word in the current phrase is equal to/less than or equal to 0-13
Number of content words before/after the current word in the current phrase is equal to/less than or equal to 0-9
Number of words from previous/next content word is equal to/less than or equal to 0-5
Number of syllables in the previous/current/next phrase is equal to/less than or equal to 0-20
Number of words in the previous/current/next phrase is equal to/less than or equal to 0-15
Forward/backward position of the current phrase in the utterance is equal to/less than or equal to 0-4
Number of syllables in the utterance is equal to/less than or equal to 0-28
Number of words in the utterance is equal to less than or equal to 0-13
Number of phrases in the utterance is equal to/less than or equal to 0-4

Figure: Non-segmental features.


Dense (1)
Dense (ch)
Dropout
ReLU
Weight Normalization
Dilated Causal Convolution
1 × 1 Convolution

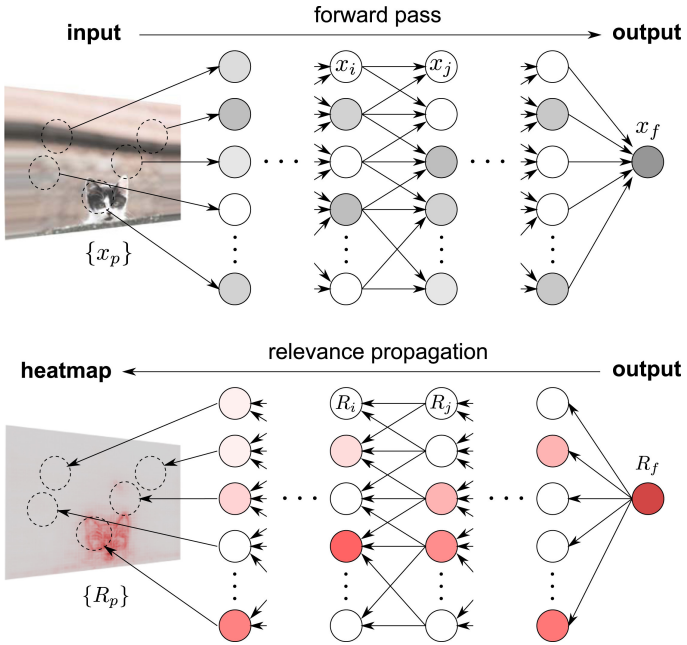


Figure: Computational flow of deep Taylor decomposition. (Adopted from Montavon 2017).

The relevance in this framework can be defined as:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k \tag{3}$$