

# 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 ( $F_0$ ), and a grammatical (phonological) aspect."* (Féry 2016)

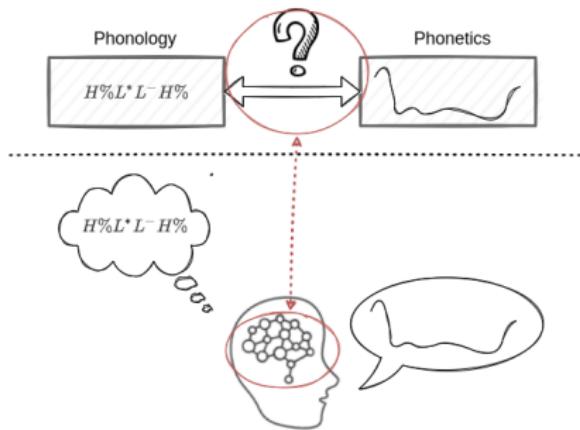
# Intonation

## Definition

*"The tonal structure of speech expressed by the melody produced by our larynx. It has a phonetic aspect, the fundamental frequency ( $F_0$ ), 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  $F_0$ .
- ✓ How linguistic features of an utterance influence its  $F_0$  contours.
- ✓ Need for a physicalist (neurobiological) model.
- ✓ Modern statistical-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 ( $F_0$  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  $F_0$  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 ( $F_0$  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  $F_0$  contours of an utterance.

## Hypotheses

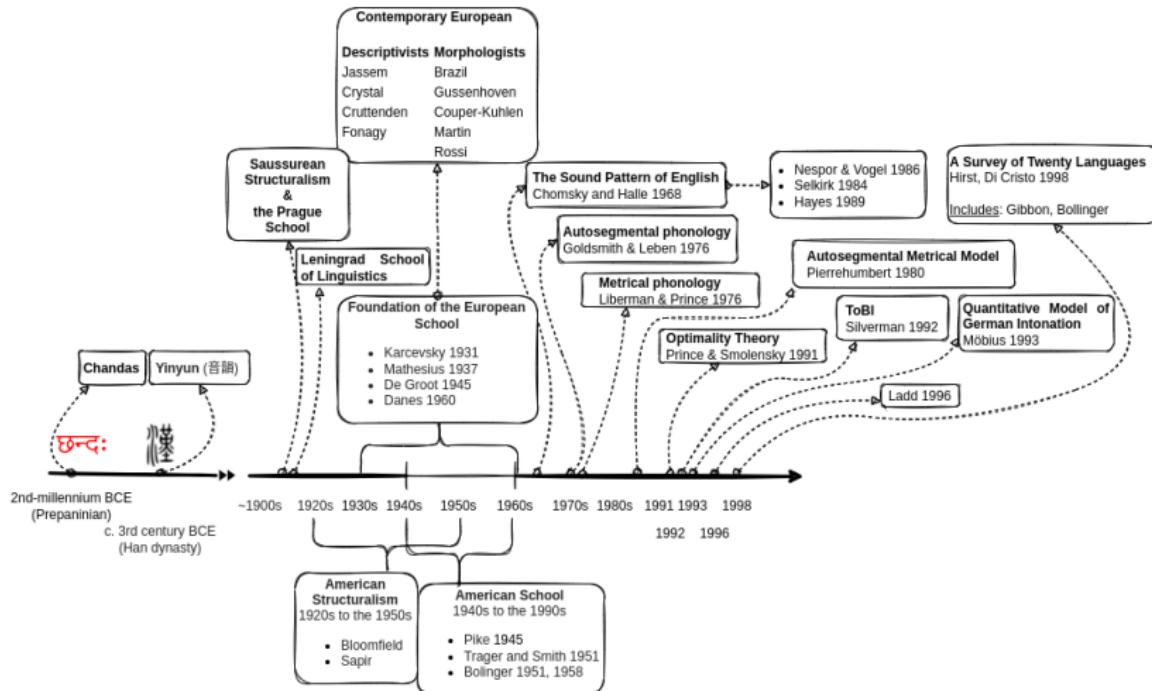
**HYPOTHESIS 1:** *The continuous  $F_0$  contours of an utterance emerge from its discrete linguistic features through a series of successive probabilistic mappings into intermediate latent representations.*

**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.*

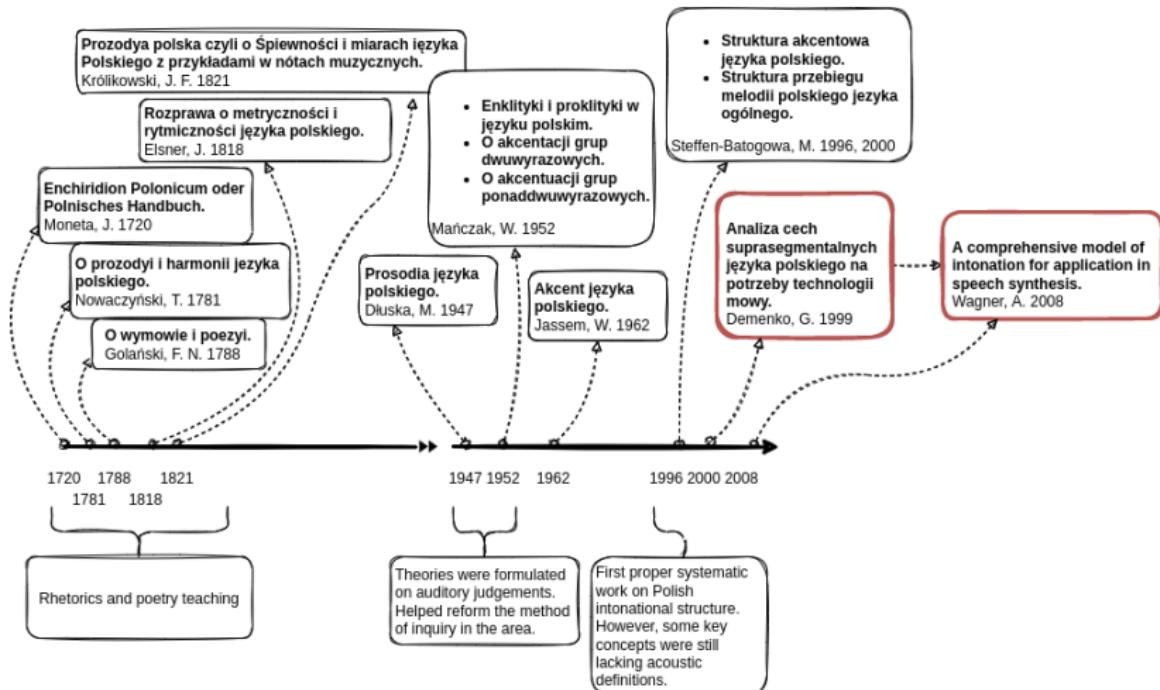
**HYPOTHESIS 3:** *The set of shallow linguistic features used in this thesis provides information which is sufficient for synthesis of natural sounding intonation in the context of statistical-parametric speech synthesis.*

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

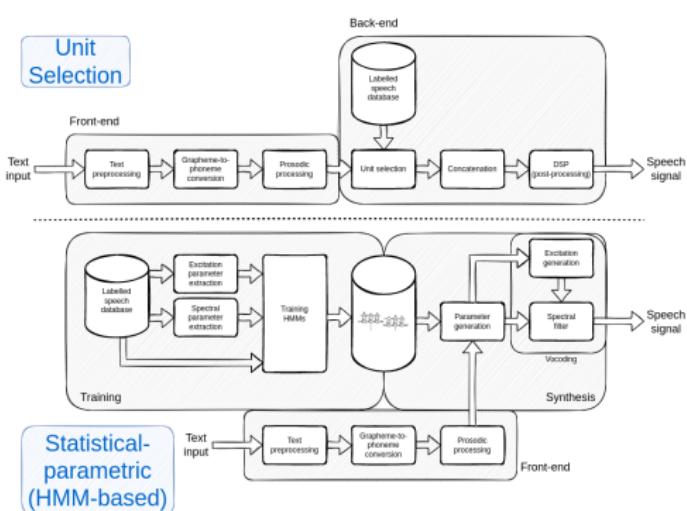
# Background



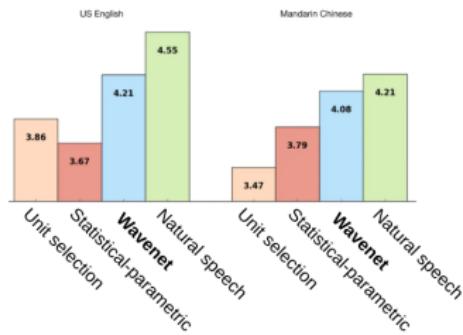
# Background



# Speech synthesis

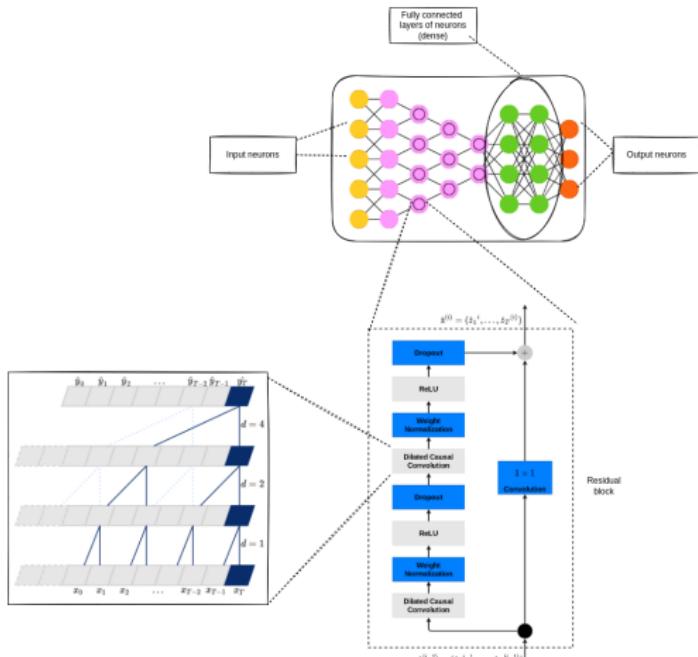


# Speech synthesis - Wavenet



**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 image recognition where they excel.



# Visual processing in the human brain

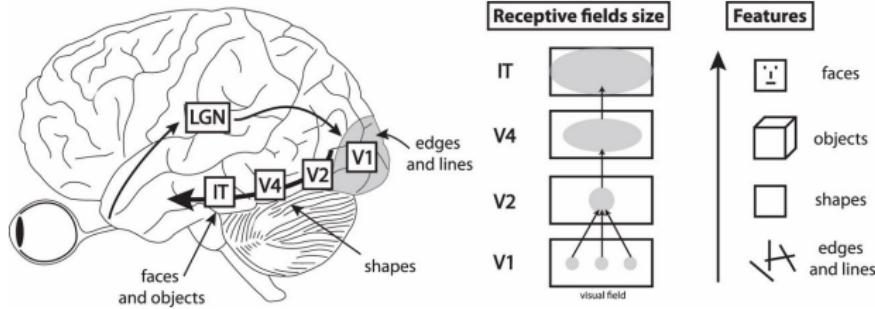


Figure: Hierarchical, feedforward visual processing in human brain. (Adopted from Manassi et al. 2013)

## CNN feature maps

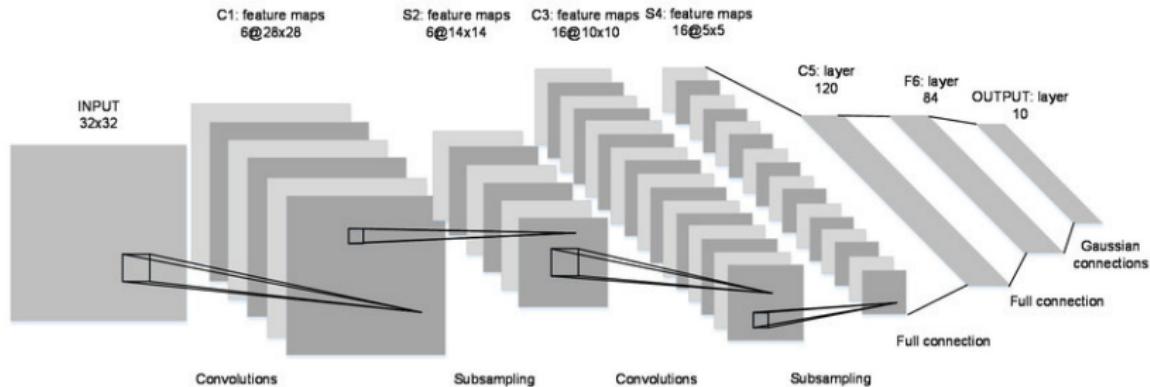
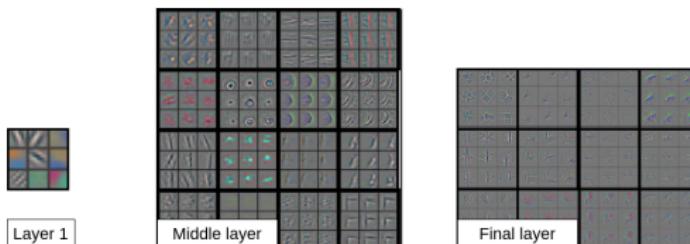
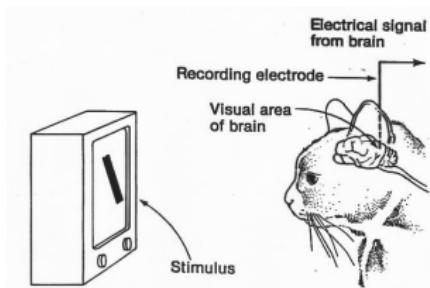


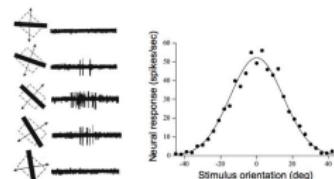
Figure: Image recognition Convolutional Neural Network (LeNet-5). (Adopted from LeCun et al. 1989)



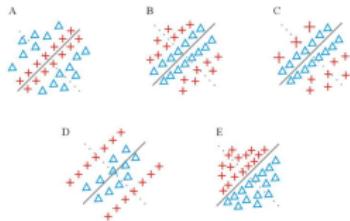
# Simple and complex cells in the visual cortex



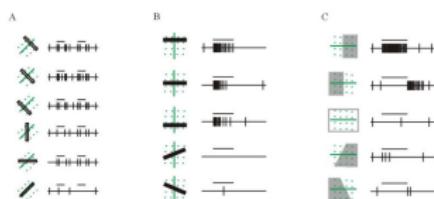
**Figure:** Famous Hubel and Wiesel cat experiment.  
(adopted from Hubel and Wiesel 1959).



**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

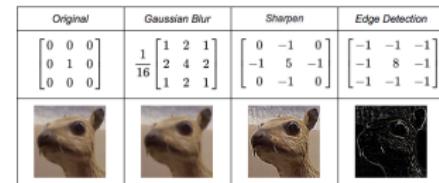
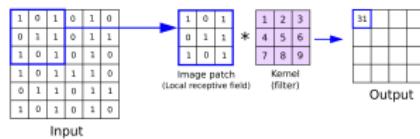


Figure: Example of a 2-dimensional matrix convolution.

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

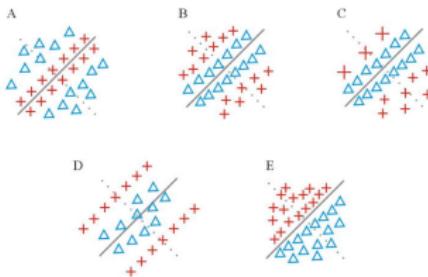


Figure: Simple receptive fields. (adopted from Hubel and Wiesel 1962).

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

# Neurobiological foundations of CNNs

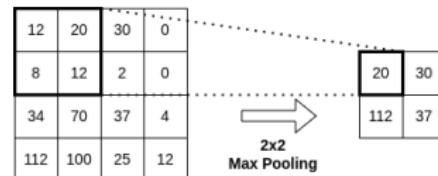


Figure: Example of a  $2 \times 2$  max pooling matrix operation.

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

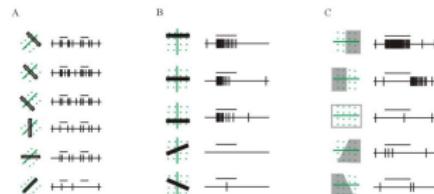
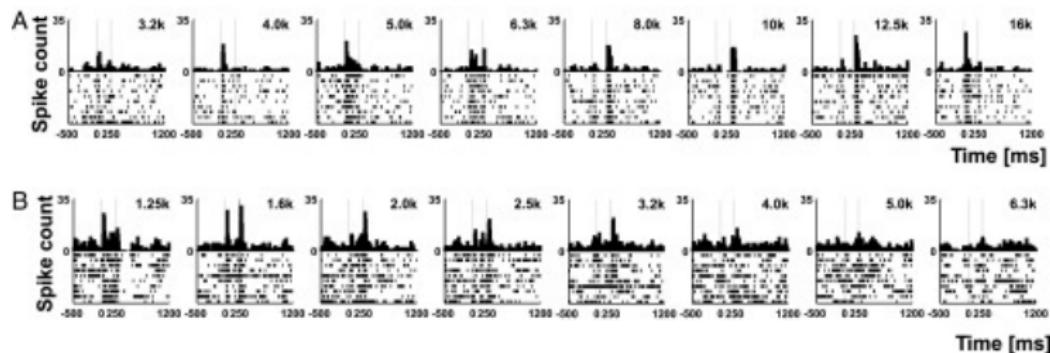


Figure: Three different types of complex receptive fields.  
(adopted from Hubel and Wiesel 1962).

# 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

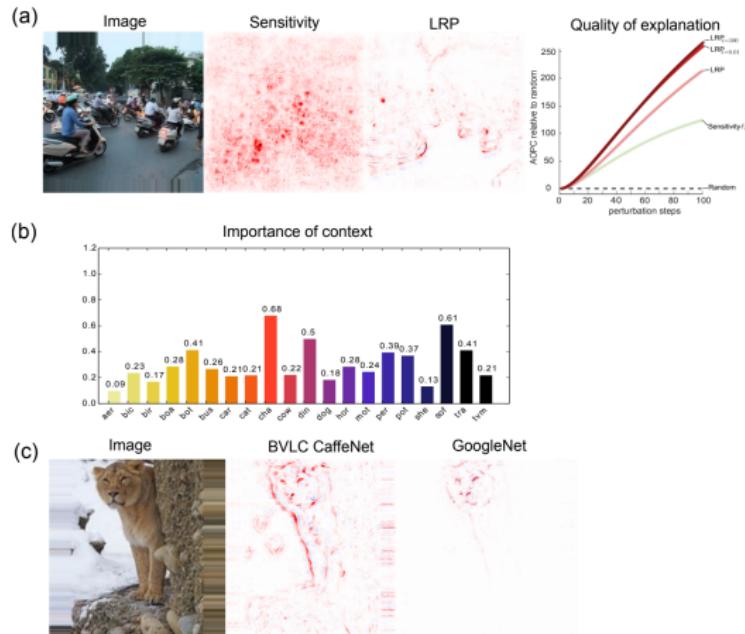
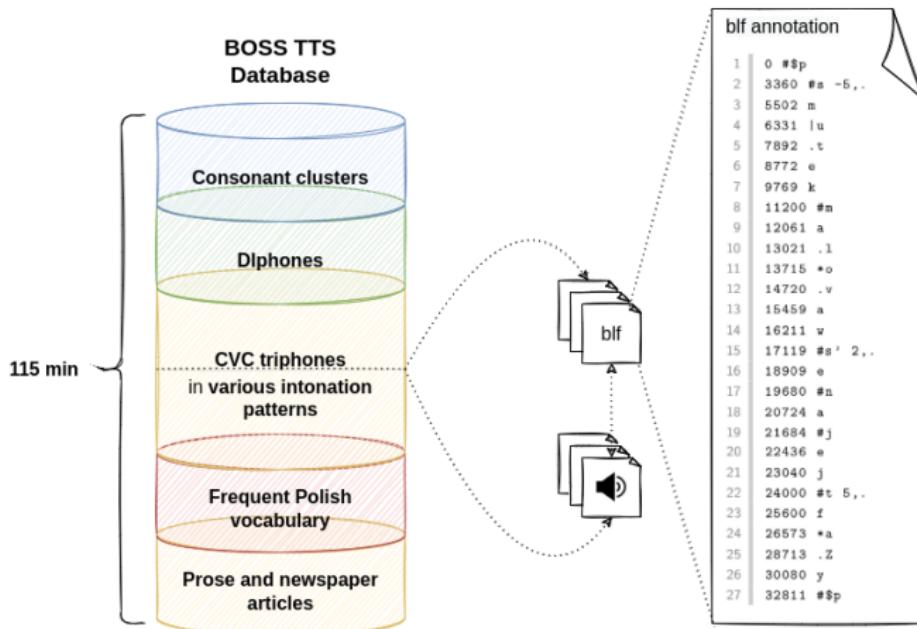


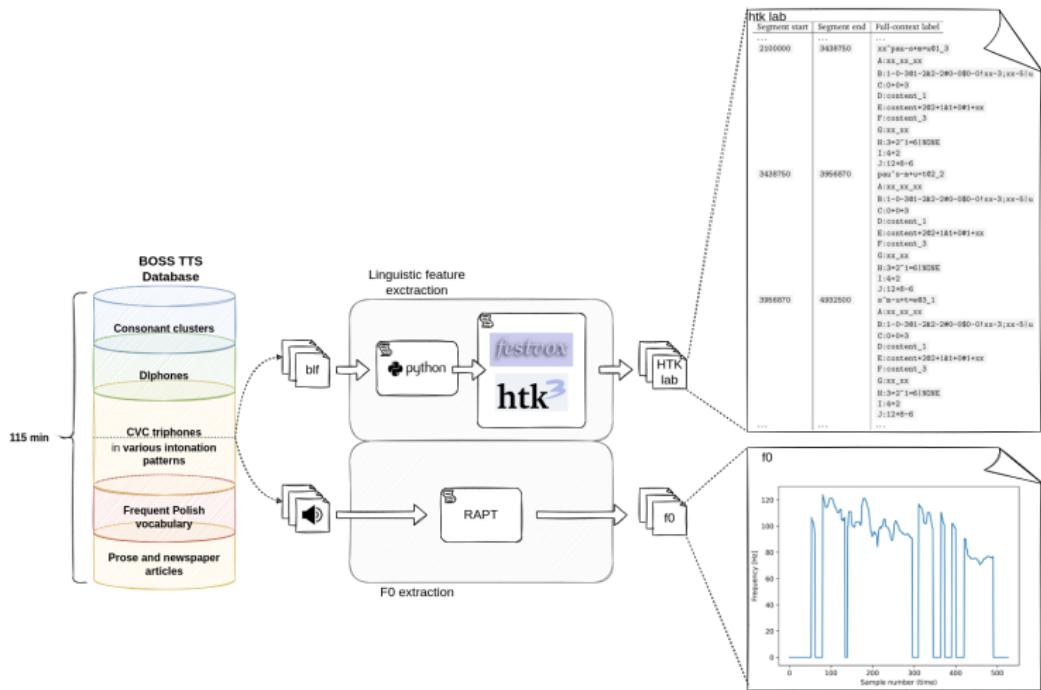
Figure: Results of sensitivity-based and relevance-based explainability methods. (Based on Samek 2016)

## Dataset

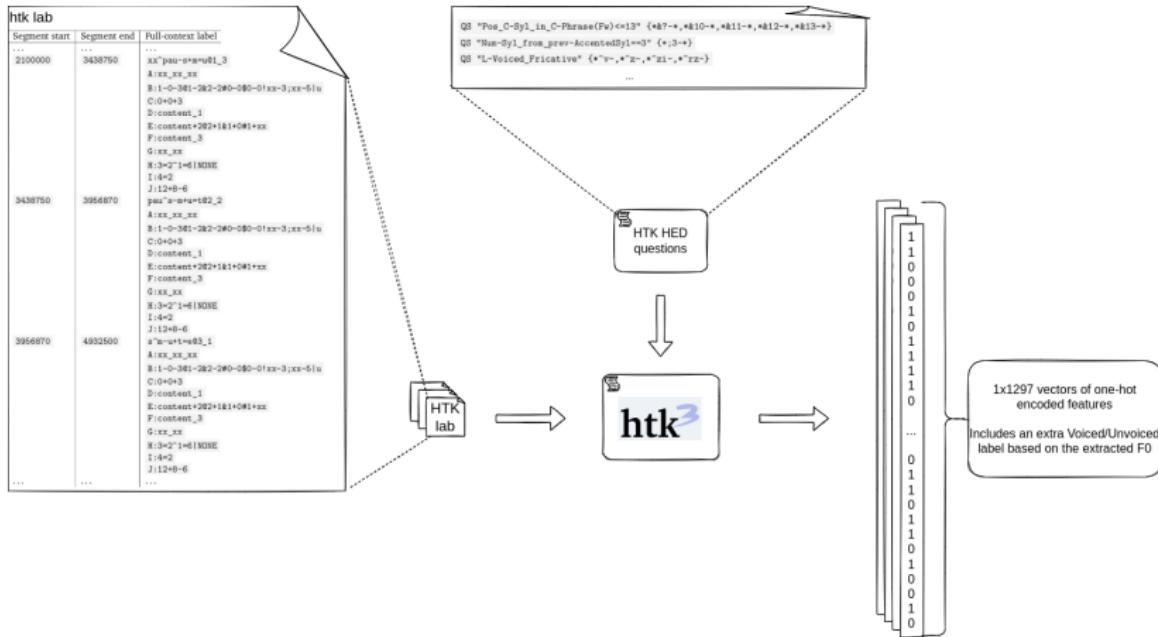


**Figure:** Speech corpus built originally for the purpose of the Polish BOSS unit selection synthesizer (Demenko, Bachan, Möbius 2008; Demenko, Klessa, Szymański, Breuer, Hess 2010).

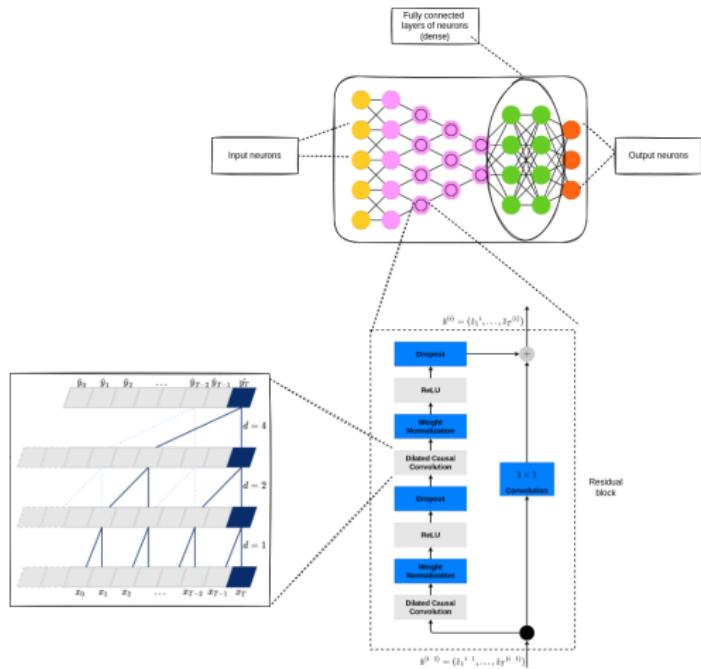
## Data preprocessing



## Feature extraction



## Model implementation



## TCN parameters

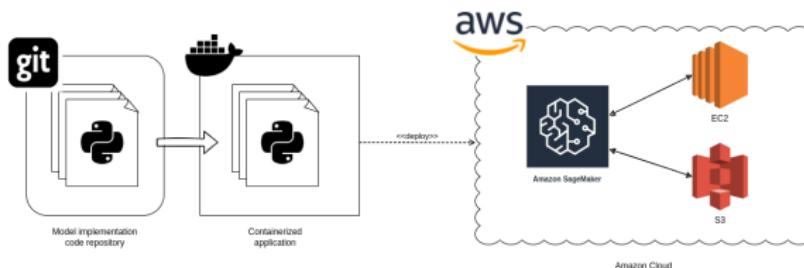
- 6 residual layers,
  - 64 convolution filters of length 2
  - $2^n$  dilations ( $[1, 2, 4, 8, 16, 32]$ )
  - 2 final fully connected dense layers  
(ReLU activation; sizes 64 and 1)

### Complete code repository

✓ [https://github.com/mrsslacklines/intonation synthesis](https://github.com/mrsslacklines/intonation-synthesis)

[github.com/philipperemy/keras-tcn](https://github.com/philipperemy/keras-tcn)

## Model training infrastructure



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

### Memory requirements

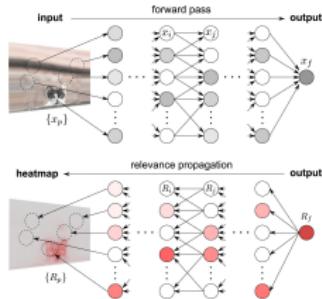
A single batch of data, which is a  $64 \times 1900 \times 1297$  vector of 8-byte boolean values, occupies 1.2617216 gigabytes of memory, and the model comprises of a total of 449,409 parameters (446,337 trainable and 3,072 non-trainable).



### Training parameters:

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

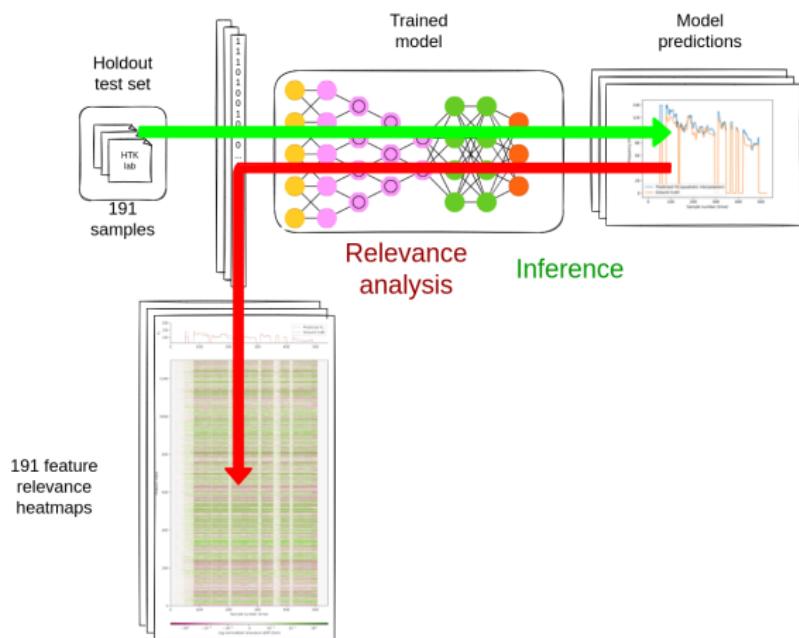
# $F_0$ inference and feature relevance analysis



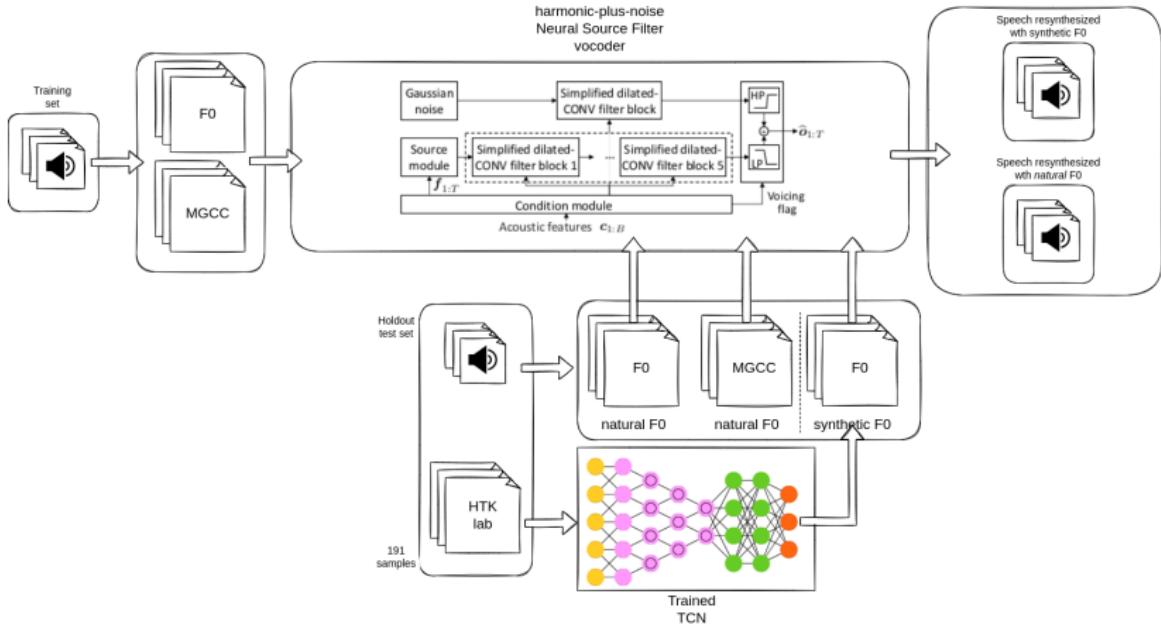
**Figure:** Computational flow of deep Taylor decomposition.  
(Adopted from Montavon 2017).

INNvestigate Neural Networks!  
(Alber et al. 2019)

[github.com/albermax/  
innvestigate](https://github.com/albermax/innvestigate)

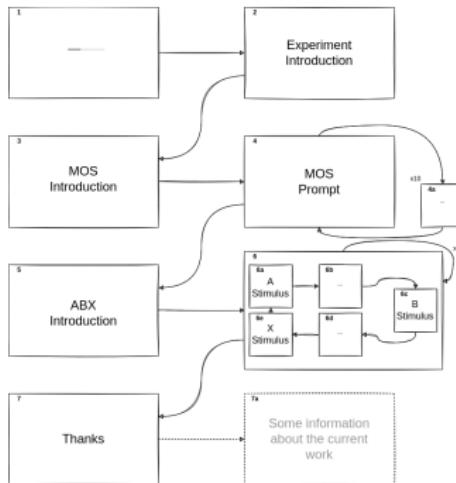
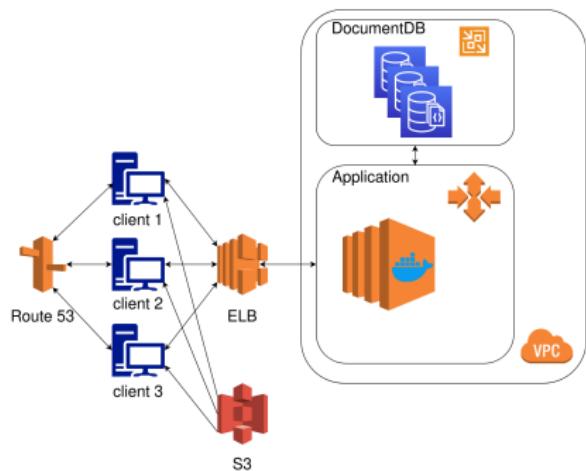


# Resynthesis



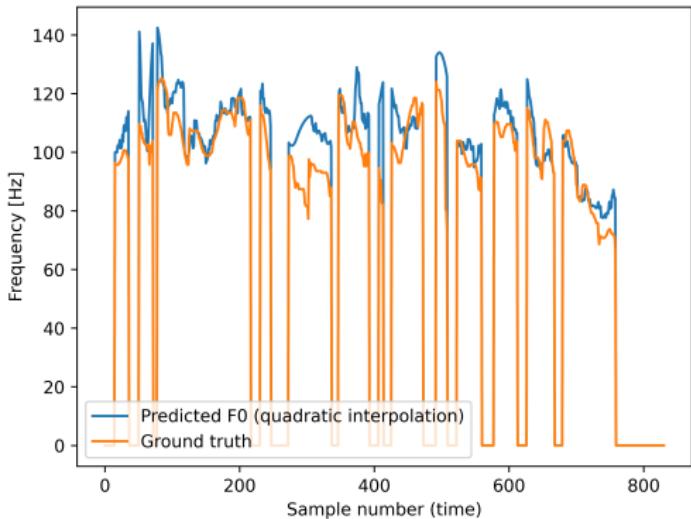
Neural Source Filter Vocoder from (X. Wang et al. 2019)

# Perceptual evaluation experiment



**Available at:**  
[fonetyka.cudaniewidły.org/experiment](http://fonetyka.cudaniewidły.org/experiment)  
**Code at:**  
[github.com/mrslacklines/listening\\_experiments](https://github.com/mrslacklines/listening_experiments)

## $F_0$ 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*).

## $F_0$ inference results

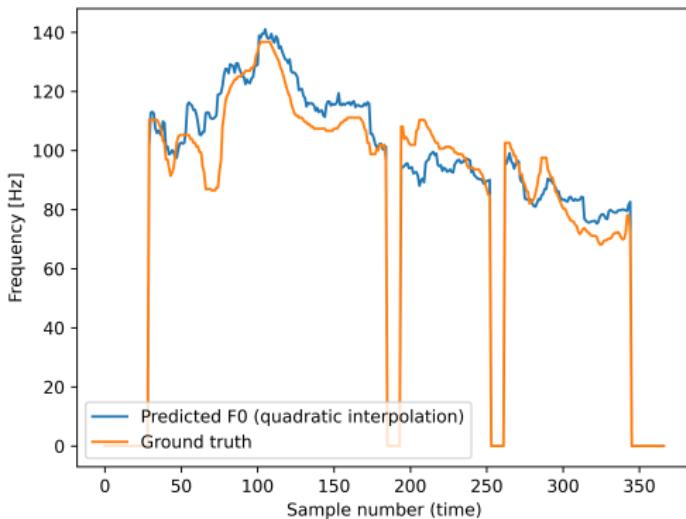


Figure: Result of  $F_0$  prediction for "Powodzenie nie jest gwarantowane" (*Success is not guaranteed*).

## $F_0$ inference results

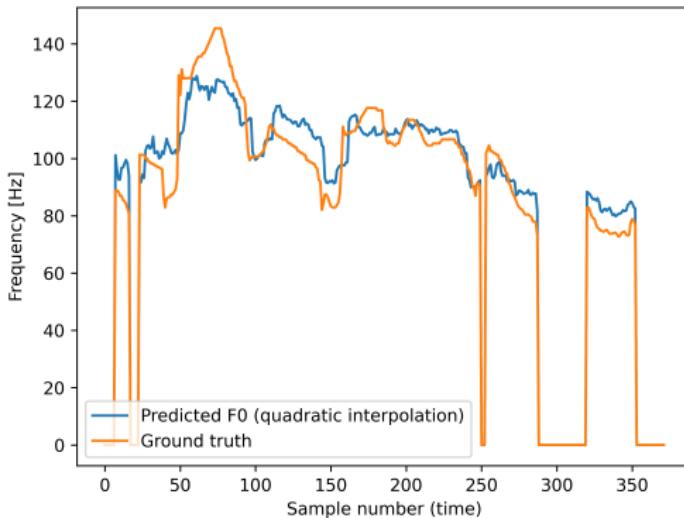


Figure: Result of  $F_0$  prediction for "Gaduła była bardzo nieznośna" (*Gabby was very annoying*.).

## $F_0$ inference results

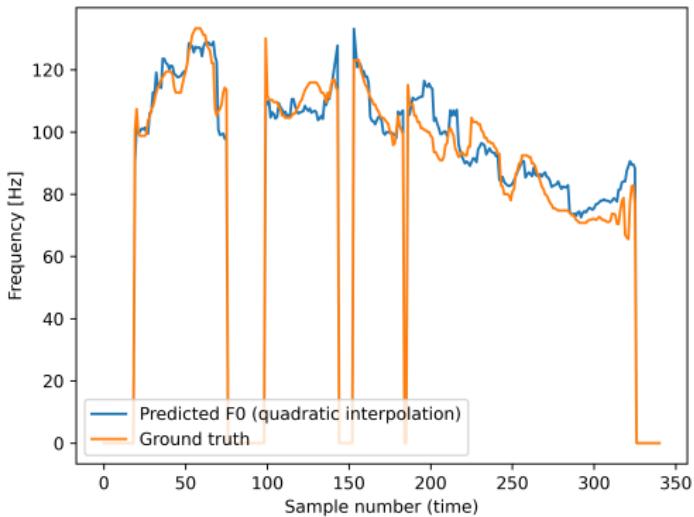


Figure: Result of  $F_0$  prediction for "Może przyniosą też gorzałę" (Maybe they will bring booze too.).

## $F_0$ inference results

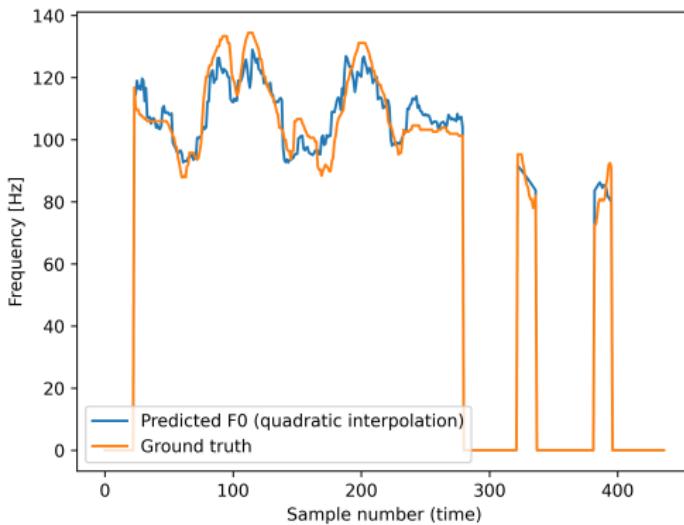
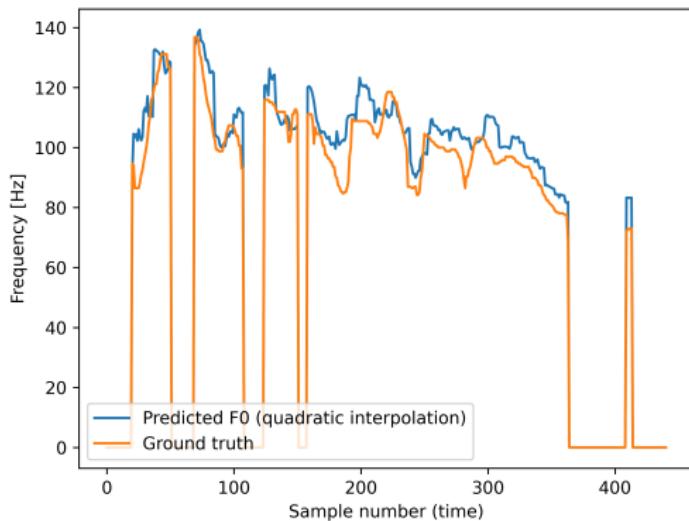


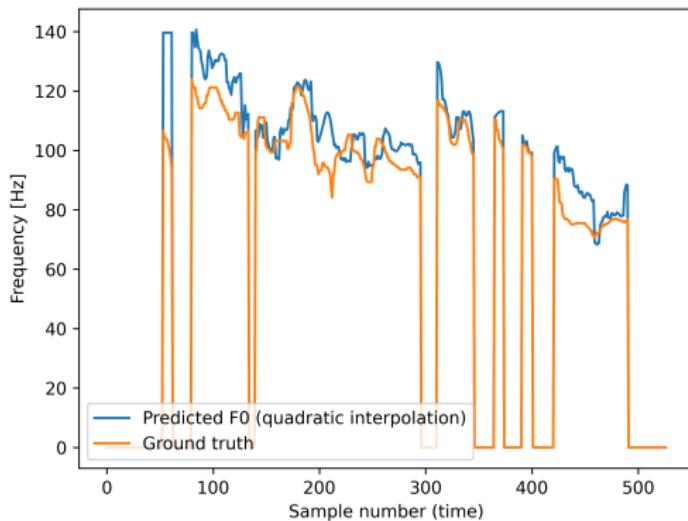
Figure: Result of  $F_0$  prediction for "To jest ważna godzina dla nas wszystkich" (*This is an important hour for all of us*).

## $F_0$ 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*).

## $F_0$ inference results



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

## $F_0$ inference results

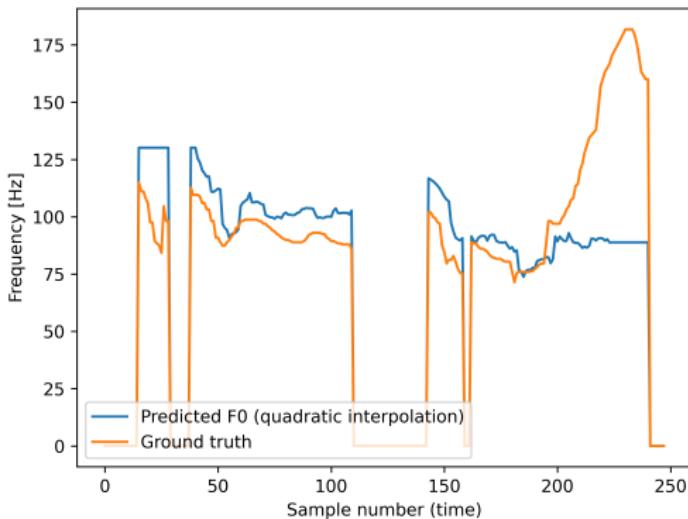


Figure: Result of  $F_0$  prediction for "Czy to był łatwy dobór?" (*Was it an easy choice?*).

## $F_0$ inference results

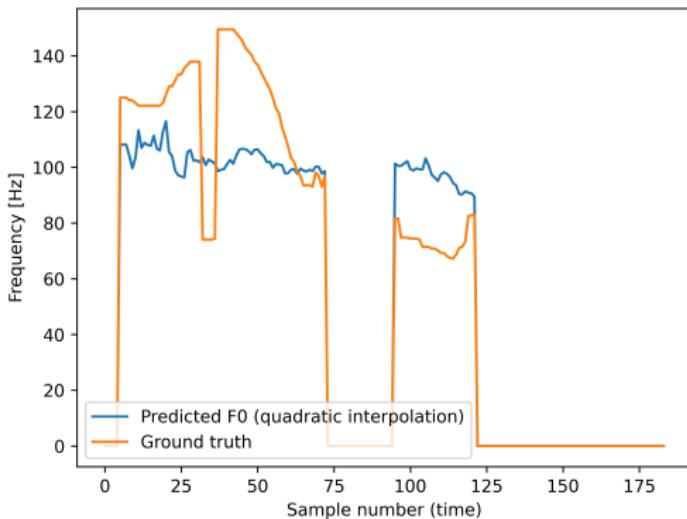


Figure: Result of  $F_0$  prediction for "To Majka" (*This is Majka*).

## $F_0$ inference results

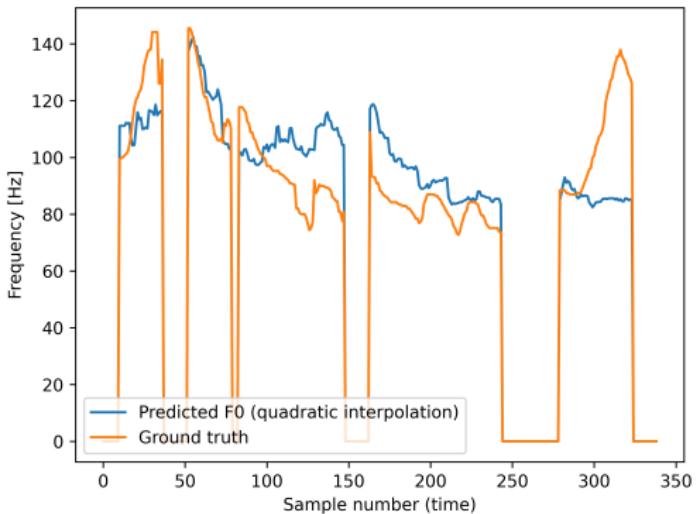


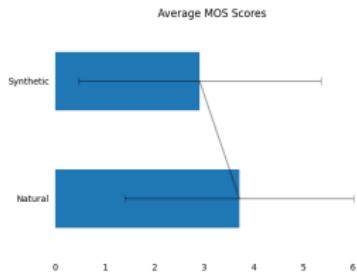
Figure: Result of  $F_0$  prediction for "Na czym polega kandyzacja?" (*How does candying work?*).

## Objective evaluation results

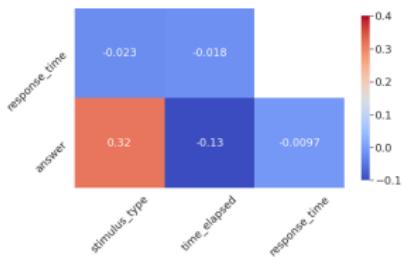
MAE	0.015382
MSE	0.002495
NRMSE	0.009534
RMSE	0.048128

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

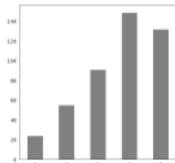
## Subjective evaluation results - MOS



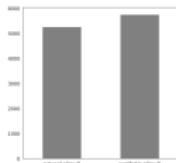
**Figure:** Mean Opinion Score-based evaluation results.



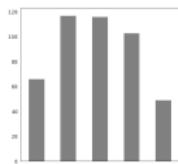
**Figure: Mean Opinion Score-based evaluation parameters correlation matrix.**



**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.



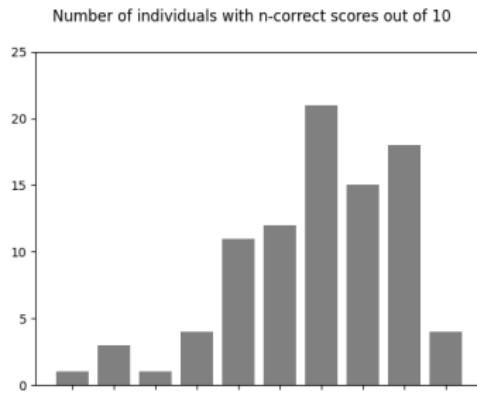
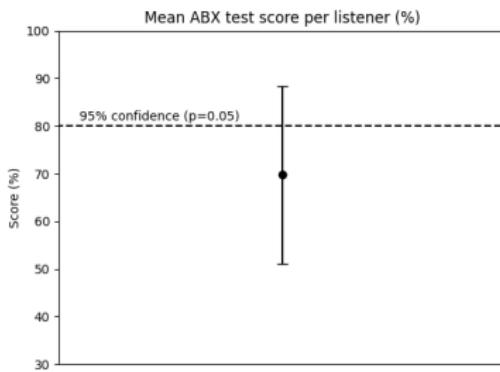
**Figure:** Mean Opinion Score-based evaluation total numbers of specific scores for synthetic stimuli.

## Subjective evaluation results – ABX

### $\tilde{\chi}^2$ test

**NULL HYPOTHESIS ( $H_0$ ):** *There are no perceptually significant differences between resynthesized recordings with synthetic and natural  $F_0$ .*

**HYPOTHESIS ( $H_a$ ):** *There are perceptually significant differences between resynthesized recordings with synthetic and natural  $F_0$ .*



## Subjective evaluation results – ABX

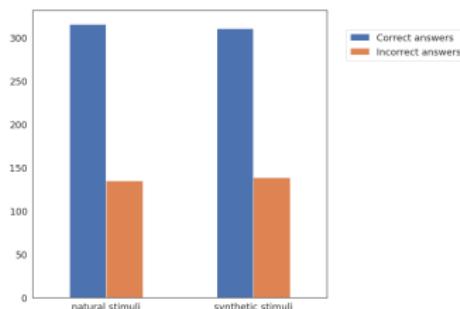


Figure: ABX experiment number of correct and incorrect answers in case of natural and synthetic stimuli.

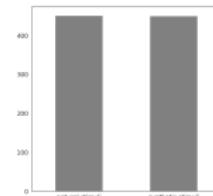


Figure: ABX experiment mean response times for synthetic and natural stimuli.

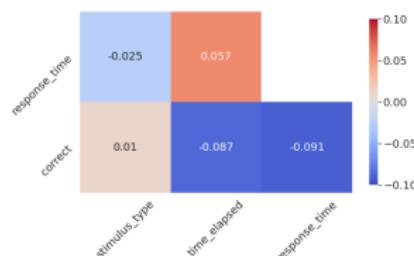
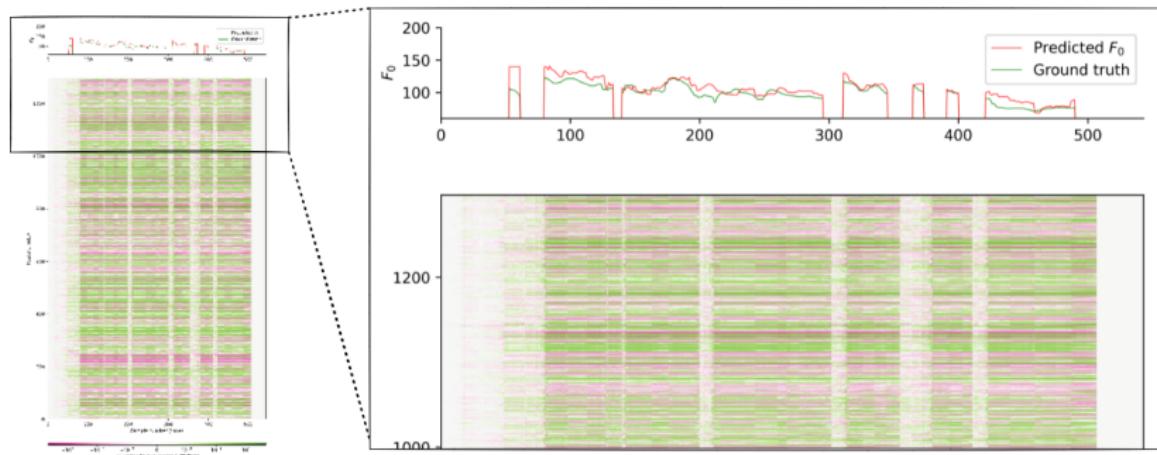


Figure: ABX experiment parameters correlation matrix.

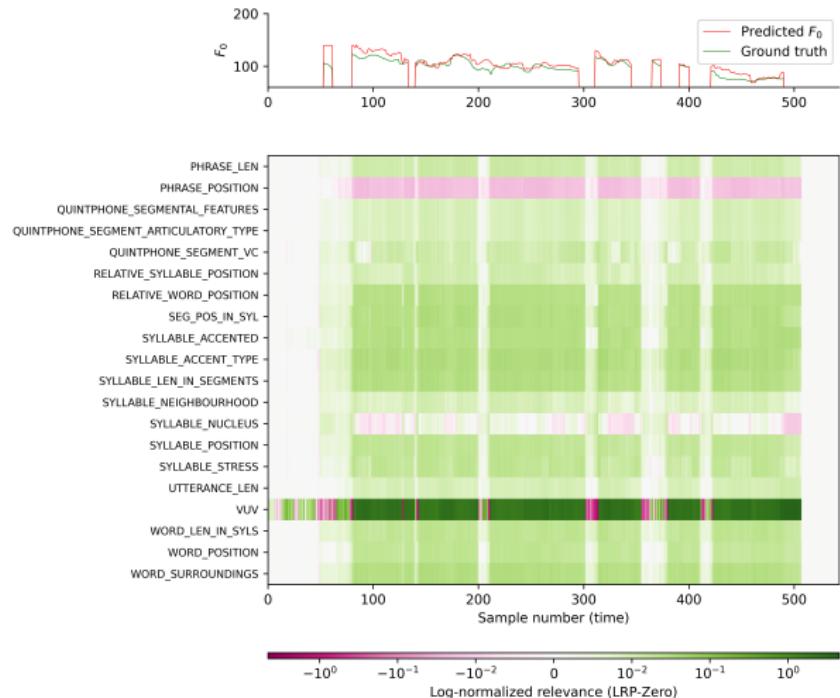
# Feature relevance analysis results



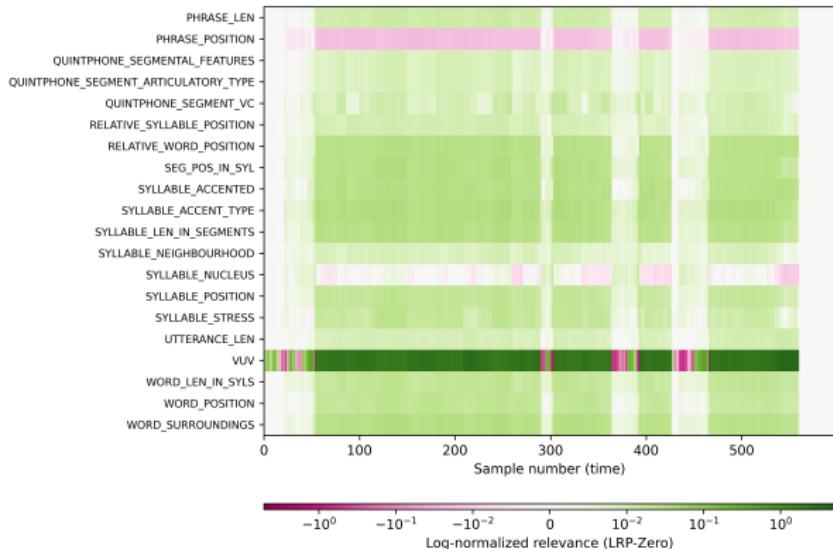
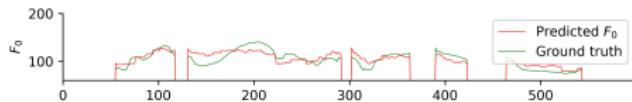
**Figure:** Fundamental frequency predictions for "Słyszałam odgłos zbliżającego się pociągu" (*I heard the sound of an approaching train*) aligned with feature relevance heatmap.

## Feature relevance analysis results

**Figure:** Fundamental frequency predictions for "Słyszałam odgłos zbliżającego się pociągu" (*I heard the sound of an approaching train*) aligned with relevance heatmap for general feature groups.



## Feature relevance analysis results

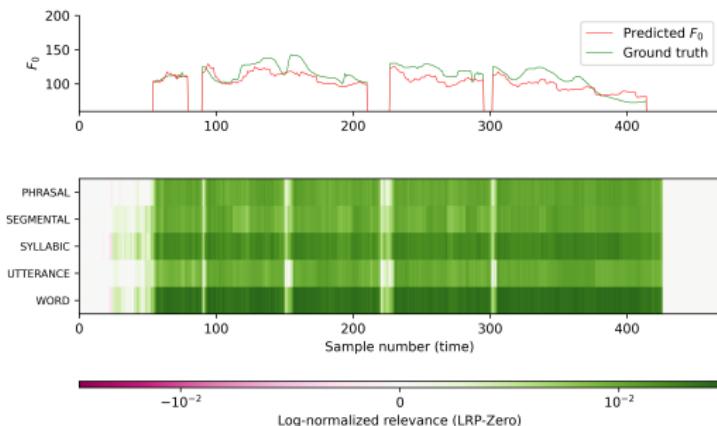


**Figure:** Fundamental frequency predictions for "Waluta jeny - w języku greckim - jest cenna" (*The currency yen, in Greek, is valuable*) aligned with relevance heatmap for general feature groups.



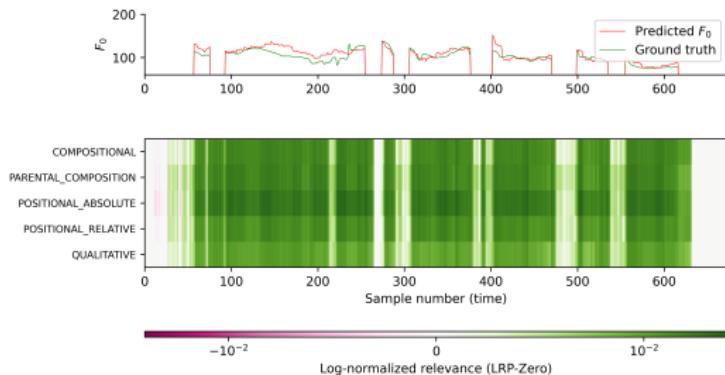
# Feature relevance analysis results

**Figure:** Fundamental frequency predictions for "Musisz dojrzeć do tego, by to zrozumieć" (*You have to grow up to understand it*) aligned with relevance heatmap for features grouped by the level of utterance.

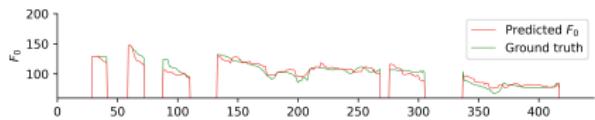


## Feature relevance analysis results

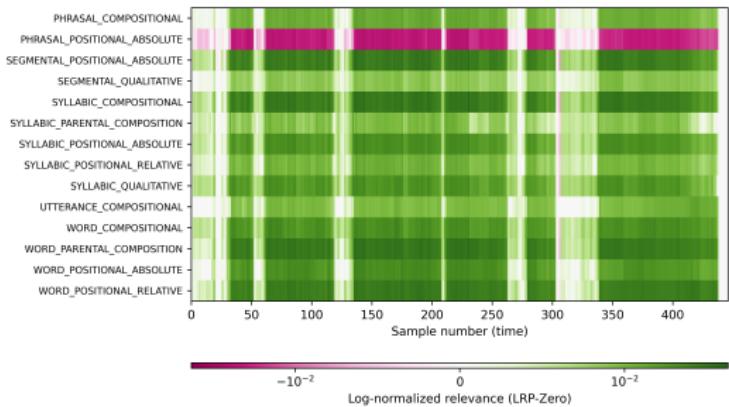
**Figure:** Fundamental frequency predictions for "Przepłynęłam na grzbiecie siedemnaście długości basenu" (*I swam seventeen pool lengths on my back.*) aligned with relevance heatmap for features grouped by the type of relation.



# Feature relevance analysis results

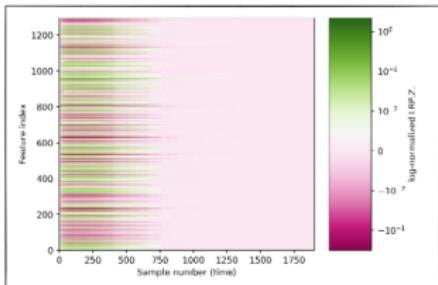


**Figure:** Fundamental frequency predictions for "Ciocia pracuje w urzędzie państwowym" (*Aunt works in a government office*) aligned with relevance heatmap for features grouped by the type of relation and the level of utterance.

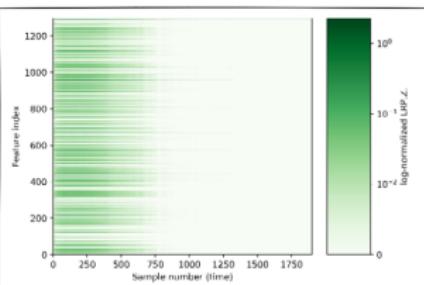
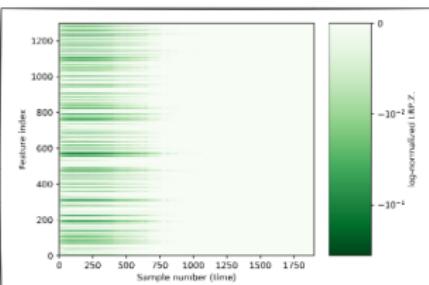
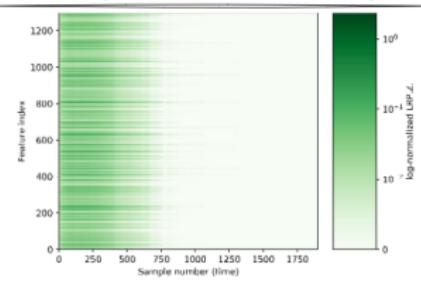


# Feature relevance analysis results

Mean



Mean (from absolute sum)

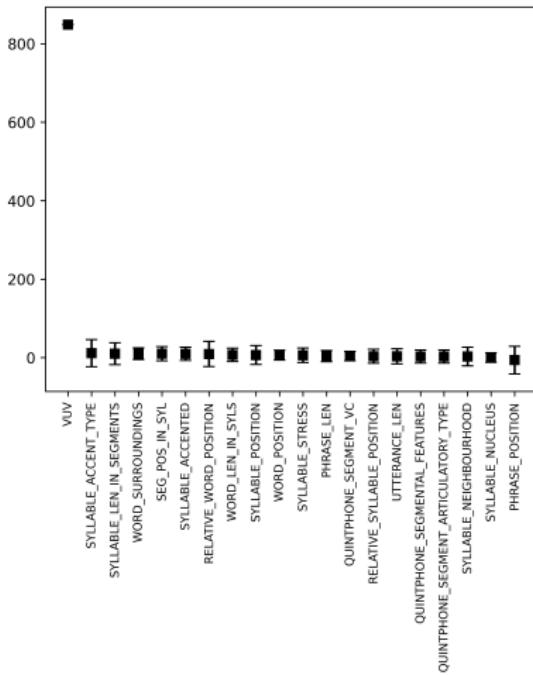


Mean (negative-only values)

Mean (positive-only values)

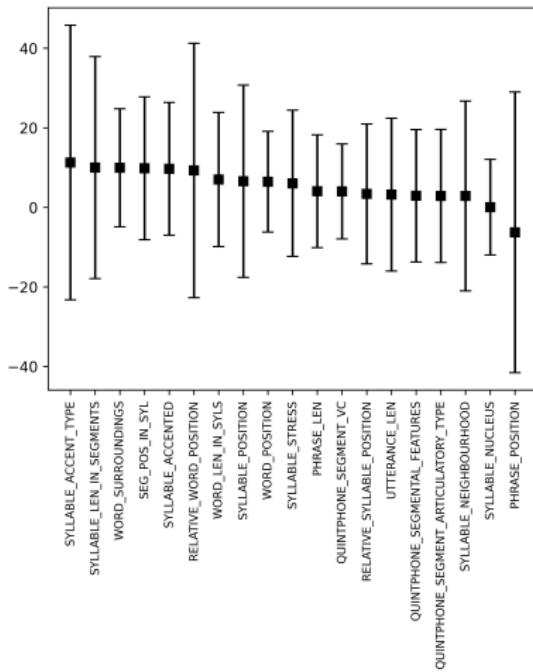
## Feature relevance analysis results

**Figure:** Regular mean relevance per feature group. The y-axis scaling due to the high relevance of V/UV renders comparatively flat plots for other features.

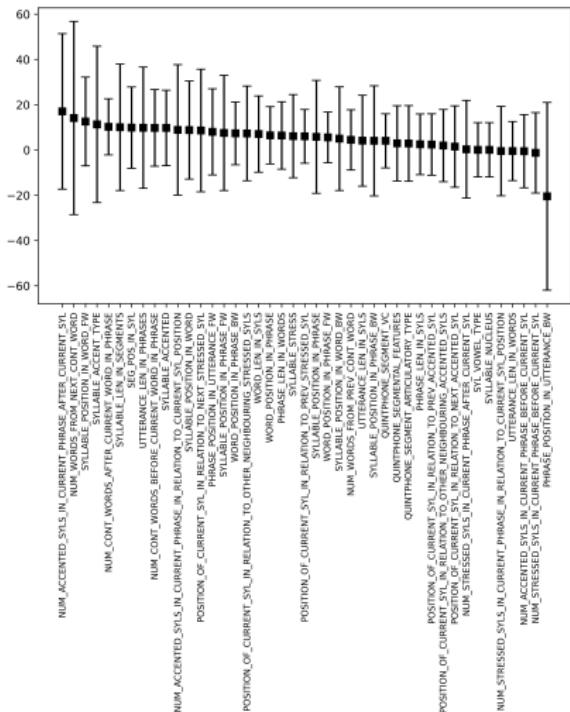


# Feature relevance analysis results

**Figure:** Regular mean relevance per feature group with the most relevant feature (V/UV) excluded.

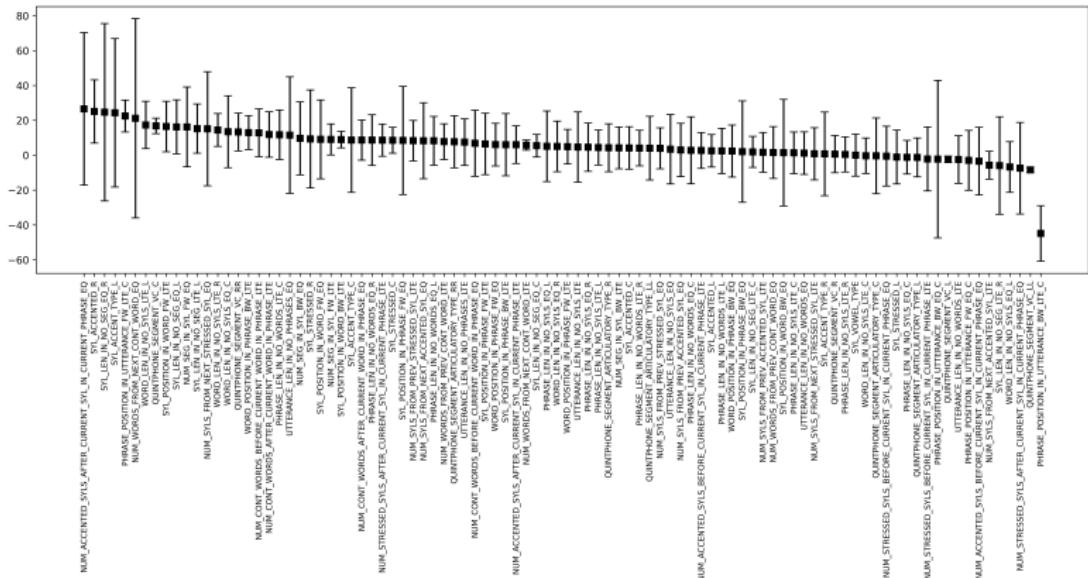


# Feature relevance analysis results



**Figure:** Regular mean relevance per feature group with the most relevant feature (V/UV) excluded. Medium granularity of feature groups.

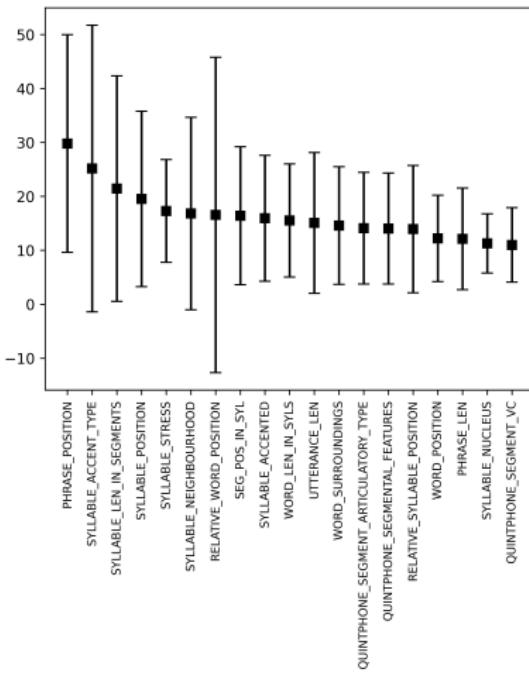
## Feature relevance analysis results



**Figure:** Regular mean relevance per feature group with the most relevant feature (V/UV) excluded. High granularity of feature groups.

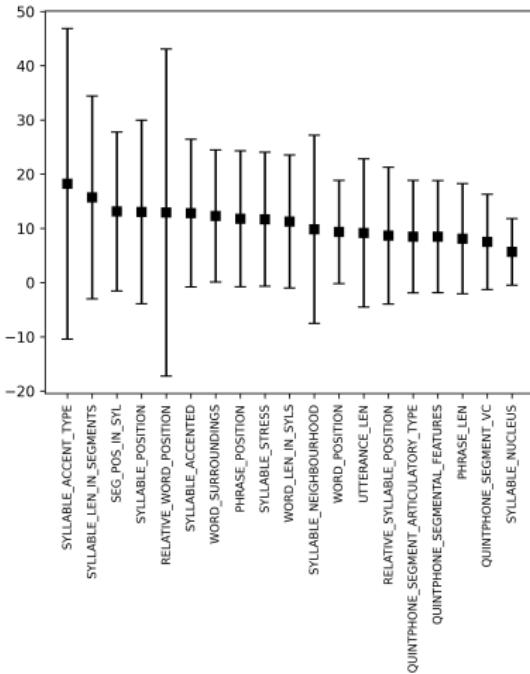
## Feature relevance analysis results

**Figure:** Absolute mean relevance per feature group with the most relevant feature (V/UV) excluded.



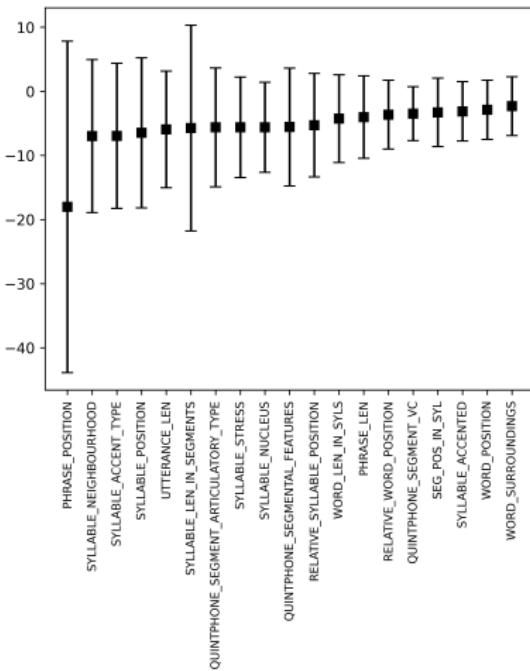
# Feature relevance analysis results

**Figure:** Positive relevance-only mean per feature group with the most relevant feature (V/UV) excluded.



## Feature relevance analysis results

**Figure:** Negative relevance-only mean per feature group with the most relevant feature (V/UV) excluded.



# Main Hypotheses

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

**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.*

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

- ✓ Generalized neural model of the mappings based on auditory cortical processing,
- ✓ Linguistically agnostic,
- ✓ Mapping model spontaneously converges towards a set of most relevant linguistic features as might be anticipated based on generally accepted linguistic theory (eg. syllabic features),
- ✓ Functional performance and naturalness of the produced  $F_0$  contours as demonstrated through the subjective evaluation procedure,
- ✓ Substantial size of results produced based on the relevance analysis of model output.

## Results

The complete results in the form of CSV files, charts, listings, etc. are available in the code repository at:  
[github.com/mrslacklines/intonation\\_synthesis](https://github.com/mrslacklines/intonation_synthesis)

## Contribution and possible applications

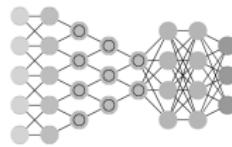
- ✓ Successful application of LRP explainability algorithm to a temporal signal modelling (speech),
- ✓ Evaluation of a biologically-motivated model of the mappings between abstract linguistic categories and the concrete contours of  $F_0$ ,
- ✓ Evidence for the effectiveness of Deep Neural Networks as explanatory scientific models and contribution to their popularization in linguistic research,
- ✓ Application in computational neuroscience research,
- ✓ Modest step towards the unification of theories of intonation,
- ✓ Methodology for determining the most relevant features that could help optimize the implementation of practical DNN-based models and the preparation of training data sets,
- ✓ Evidence that could support or undermine some of the current assumptions in phonological grammars of intonation,
- ✓ Complete open-source implementation available in public repositories including quickstart instructions.

### Code repository

[https://github.com/mrslacklines/intonation\\_synthesis](https://github.com/mrslacklines/intonation_synthesis)

## Improvements and future plans

- ✓ Size, coverage and quality of the dataset,
- ✓ Better data work, exploratory analysis (e.g. feature correlations)
- ✓ Better objective metric,
- ✓ Comparison of multiple model instances,
- ✓ Hyperparameter optimization,
- ✓ Account for the influence of the biomechanics of the articulatory tract on the final contour of the intonation,
- ✓ Measure relevance given the accuracy of the prediction,
- ✓ Redesign the model based on a single coherent neural theory and provide a better justification of the model architecture,
- ✓ Run experiments on features from other levels of language (such as semantics, syntax, pragmatics) or for expressive speech,
- ✓ Test other variants of the LRP algorithm,
- ✓ Verify the relevance of features with an ablation study,
- ✓ Design better subjective evaluation experiment,
- ✓ Analyze the intermediate latent features extracted by the deep convolutional filters.



Thank you!

## Speech synthesis - Wavenet

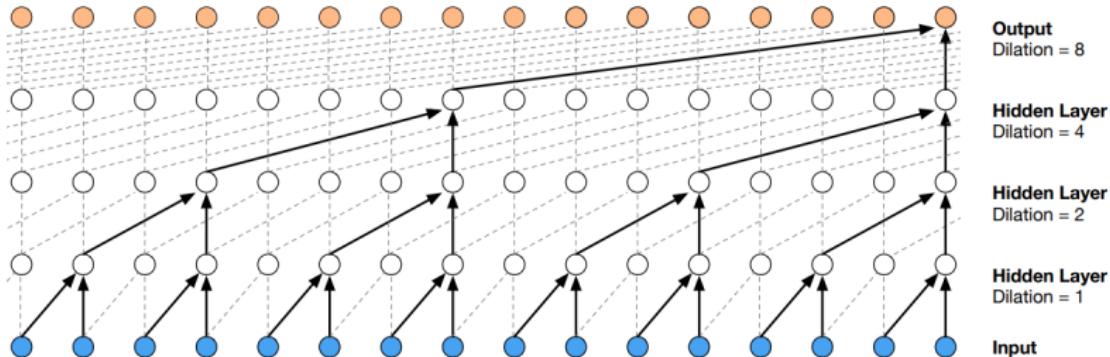
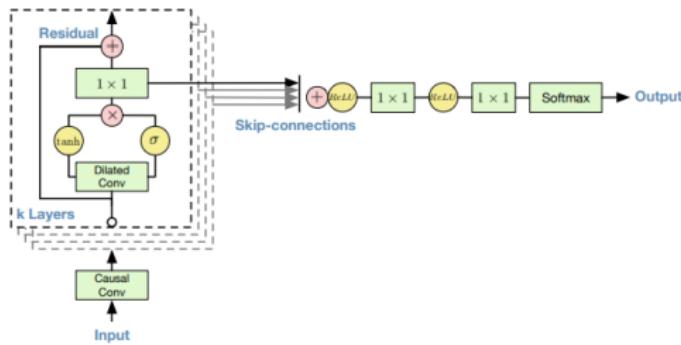


Figure: Dilated causal convolutions. (Adopted from the original WaveNet paper).

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

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

# Speech synthesis - Wavenet



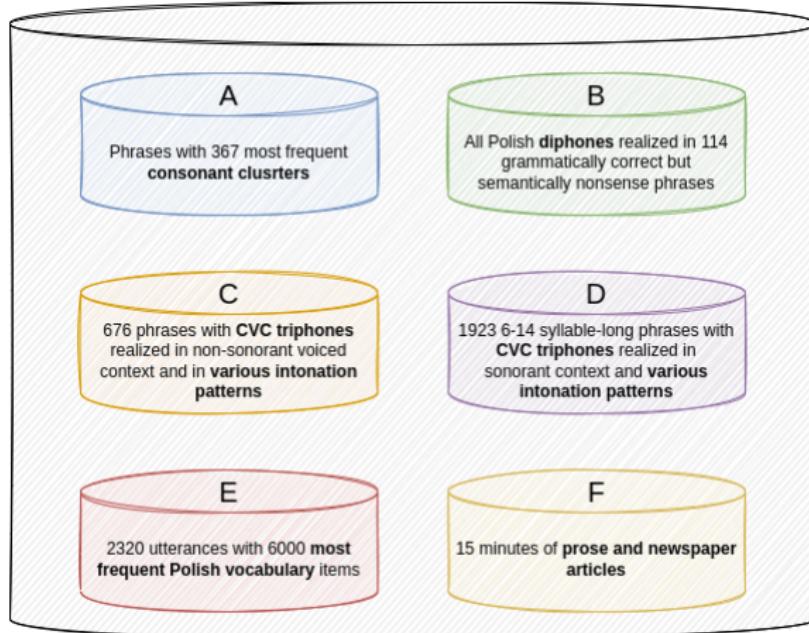
**Figure:** Residual and skip connections from a stack of  $k$  gated convolutional layers (Adopted from the original WaveNet paper).

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.

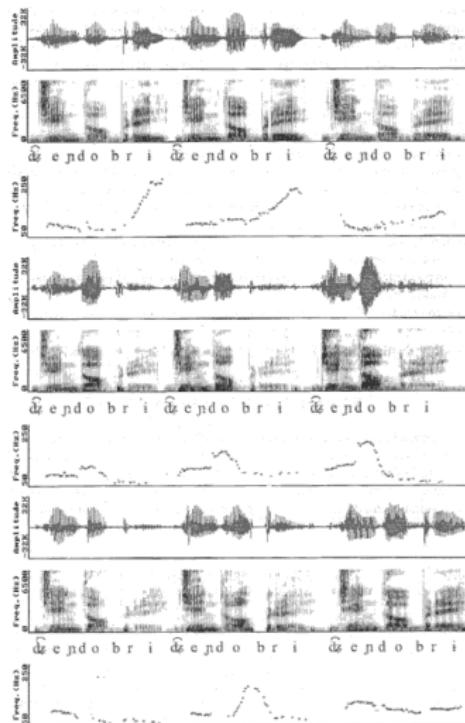


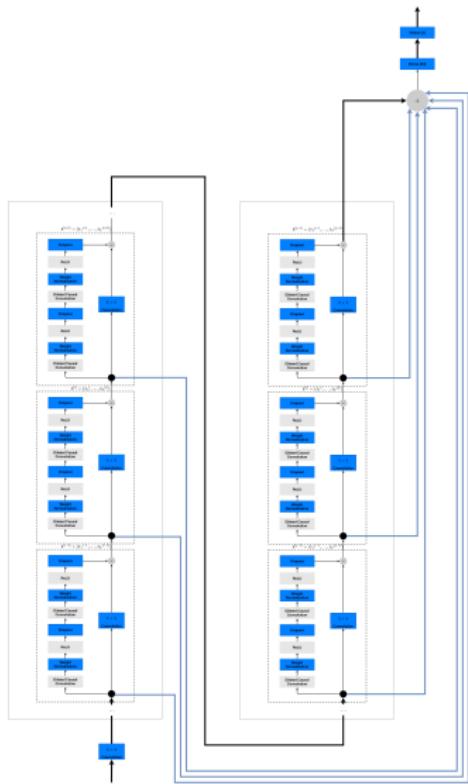
Figure: Acoustic realizations of the 9 different accents. Adopted from (Demenko 1999).

# Feature set

Question type	Segments	
Vowel	{i, y, e, a, o, u, [schwa]}	Number of preceding/succeeding segments in the previous/current/next syllable is equal to/less than or equal to 0-7
Consonant	{g, g̊, p, b, t, d, k, l, g, kl, g̊l, f, v, s, z, ñ, x, xl, m̊, m̊z, rx, x̊, c, dz, cz, drz, cl, dzl, n, n̊, ñl, ng, l, r, v, ẘ, j, jj̊}	Previous/current/next syllable is stressed
Stop	{g, g̊, p, b, t, d, k, l, g̊l}	Previous/current/next syllable has accent X (where X is one of the ToBI accents described above)
Nasal	{vn, ñn, m, n, m̊, m̊z}	Number of preceding/succeeding segments in the next syllable is equal to/less than or equal to 0-7
Frikative	{f, v, s, xl, zl, nz, rx, x̊}	Forward/backward position of the current syllable in current word is equal to/less than or equal to 0-7
Front	{e, ɔ, y, f, v, p, b, m, w, ẘ}	Forward/backward position of the current syllable in current phrase is equal to/less than or equal to 0-20
General	{[schwa], n, ñ, x, xl, s, xl, z, xl, n, r, t, d, dz, rx, zx, cz, drz, c, dn, cl, dzl}	Number of stressed syllables before/after the current syllable in current phrase is equal to/less than or equal to 0-12
Back	{o, u, x, g, xl, gl, ng, x, gx̊}	Number of accented syllables before/after the current syllable in current phrase is equal to/less than or equal to 0-12
Front Vowel	{e, ɔ, y}	Number of accented syllables before/after the current syllable in current phrase is equal to/less than or equal to 0-7
Central Vowel	{a, [schwa]}	Number of syllables from previous/next stressed syllable is equal to/less than or equal to 0-5
Back Vowel	{o, u}	Number of syllables from previous/next accented syllable is equal to/less than or equal to 0-16
High Vowel	{i, y, u}	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]
Medium Vowel	{e, ɔ}	Number of syllables in the previous/current/next word is equal to/less than or equal to 0-7
Low Vowel	{a}	Forward/backward position of the current word in the current phrase is equal to/less than or equal to 0-13
Rounded Vowel	{o, u}	Number of content words before/after the current word in the current phrase is equal to/less than or equal to 0-9
Unrounded Vowel	{a, e, i, y}	Number of words from previous/next content word is equal to/less than or equal to 0-5
Xlowl (e.g. Altvowel)	{i, y, e, a, o, u, [schwa]}	Number of syllables in the utterance is equal to/less than or equal to 0-28
Unvoiced Consonant	{g, g̊, p, t, k, kl, f, v, s, xl, x, c, cz, cl̊}	Number of words in the utterance is equal to/less than or equal to 0-15
Voiced Consonant	{b, d, g, gl, v, z, xl, rx, dz, drz, dzl, n, m, m̊, ng, l, r, v, ẘ, j, jj̊}	Forward/backward position of the current phrase in the utterance is equal to/less than or equal to 0-4
Front Consonant	{f, v, t, p, b, m, w, ẘ}	Number of words in the utterance is equal to/less than or equal to 0-13
General Consonant	{t, d, s, xl, z, xl, n, r, l, t, d, zx, rx, cz, drz, c, dz, cl, dzl}	Number of phrases in the utterance is equal to/less than or equal to 0-4
Back Consonant	{g, g̊, k, l, g̊l, ng, x̊}	
Fortis Consonant	{g, g̊, cz, f, k, p, s, nz, t, cl, c, xl̊}	
Lenis Consonant	{drz, v, g, b, rx, x, d, dn̊l, dz, gl, zl̊}	
Neither F or L	{m, n, xl, ng, l, r, v, ẘ, j, jj̊}	
Voiced Stop	{b, d, g}	
Unvoiced Stop	{p, t, k, g̊}	
Front Stop	{b, p}	
General Stop	{d, t}	
Back Stop	{g, k, g̊}	
Voiced Frikative	{v, z, xl, rx}	
Unvoiced Frikative	{f, s, xl, m̊, m̊z}	
Front Frikative	{t, v}	

Figure: Segmental features for a quintphone-wide context.

Figure: Non-segmental features.



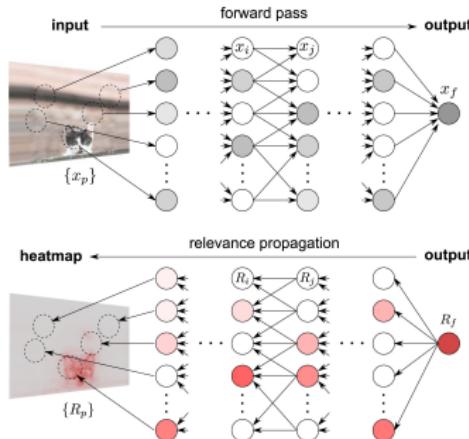


Figure: Computational flow of deep Taylor decomposition. (Adopted from Montavon 2017).

The relevance in this framework can be defined as:

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