Prak: An automatic phonetic alignment tool for Czech

Václav Hanžl[1], Adléta Hanžlová[2]

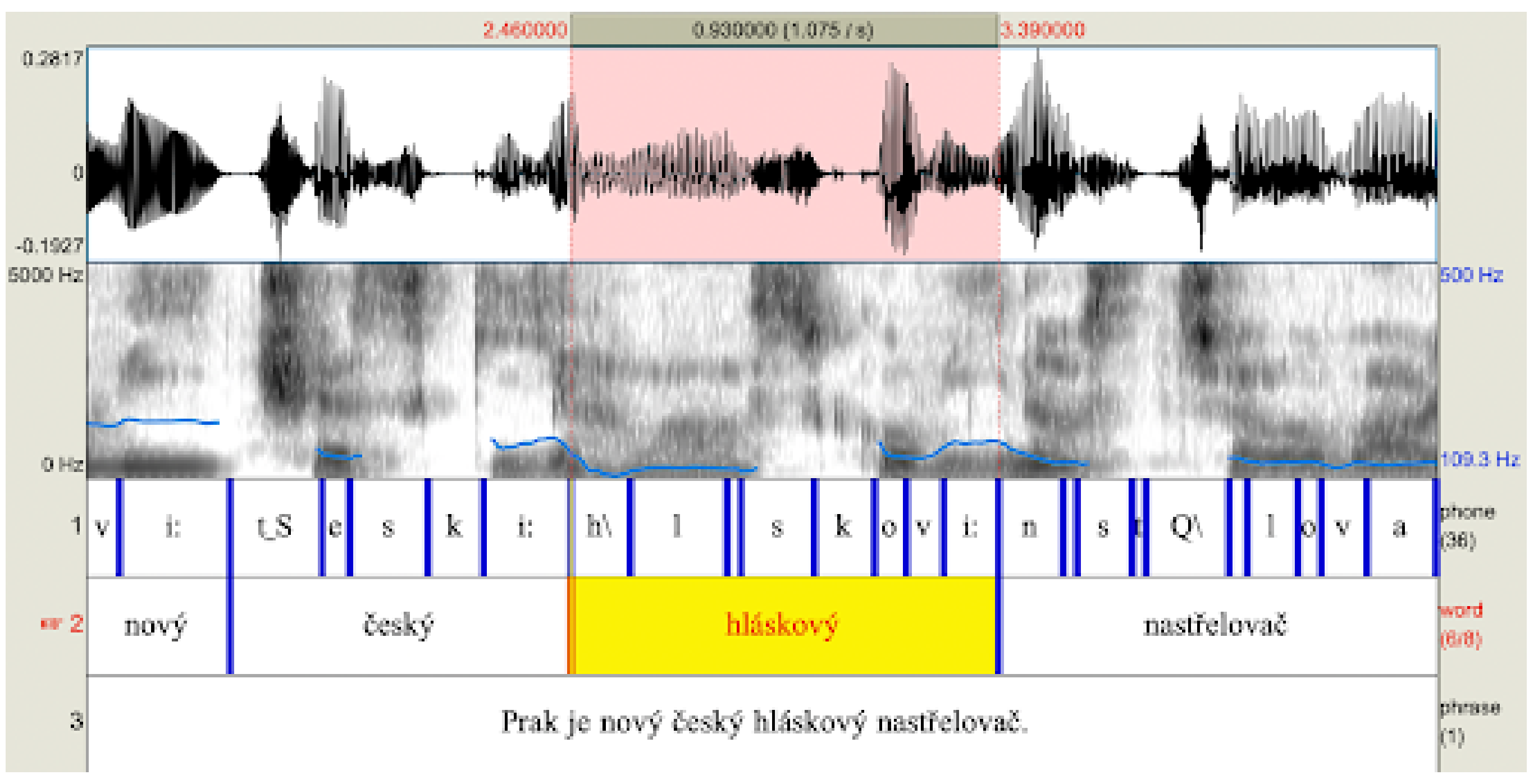
[1]Freelance researcher, [2]Institute of Phonetics, Charles University in Prague

vhanzl@gmail.com,adleta.hanzlova@gmail.com

Abstract

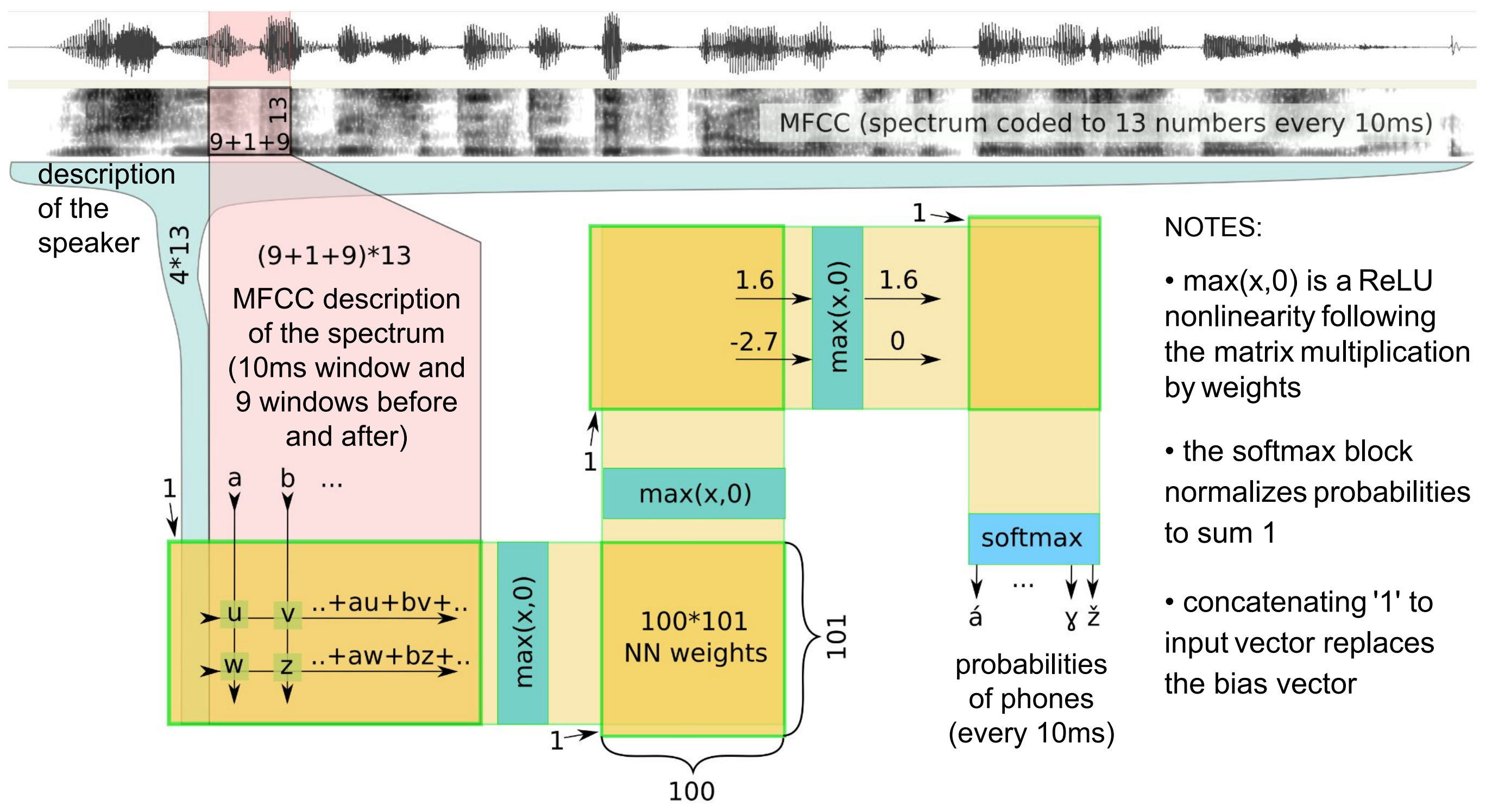
Labeling speech down to the identity and time boundaries of phones is a labor-intensive part of phonetic research. To simplify this work, we created a **free open-source tool** generating **phone sequences** from Czech text and **time-aligning** them with audio. Low architecture complexity makes the design approachable for students of phonetics. Acoustic model ReLU **NN** with 56k weights was trained using **PyTorch** on small CommonVoice data. Alignment and **variant selection** decoder is implemented in **Python** with matrix library. A Czech pronunciation generator is composed of simple rule-based blocks capturing the logic of the language where possible, allowing modification of transcription approach details. Compared to tools used until now, data preparation efficiency improved, the tool is usable on **Mac, Linux and Windows** in **Praat GUI** or command line, achieves mostly correct pronunciation variant choice including **glottal stop** detection, algorithmically captures most of Czech assimilation logic and is both didactic and practical.

Prak used in Praat GUI


2.460000
0.930000 (1.075 / s)
3.390000
0.2817
0
-0.1927
5000 Hz
0 Hz
500 Hz
109.3 Hz
1
v i: t_S e s k i: h\ l s k o v i: n s t Q\ l o v a
phone (36)
2
nový český hláskový nastřelovač
word (6/8)
3
Prak je nový český hláskový nastřelovač.
phrase (1)

Prak is fully integrated in Praat - just create a phrase tier and press the "Align by Prak" button. Behind the scenes, the Python-based aligner is run and the result is shown back in Praat. Binding script in Praat does many sanity checks and allows several modes, e.g. alignment of multiple sound+phrase couples or using common phrase for multiple sounds.

DNN Acoustic Model of Phones


MFCC (spectrum coded to 13 numbers every 10ms)
13
9+1+9
description of the speaker
4*13
(9+1+9)*13 MFCC description of the spectrum (10ms window and 9 windows before and after)
1
a
b
...
u
v
w
z
..+au+bv+..
..+aw+bz+..
max(x,0)
1
100*101 NN weights
101
100
1
max(x,0)
1.6
-2.7
max(x,0)
1.6
0
1
softmax
á
...
ɣ ž
probabilities of phones (every 10ms)
NOTES:
• max(x,0) is a ReLU nonlinearity following the matrix multiplication by weights
• the softmax block normalizes probabilities to sum 1
• concatenating '1' to input vector replaces the bias vector

For each 10ms frame, DNN AM computes probabilities of phones. Viterbi decoder then finds the best alignment for a sausage graph from the pronunciation generator.

Download Prak

https://github.com/vaclavhanzl/prak

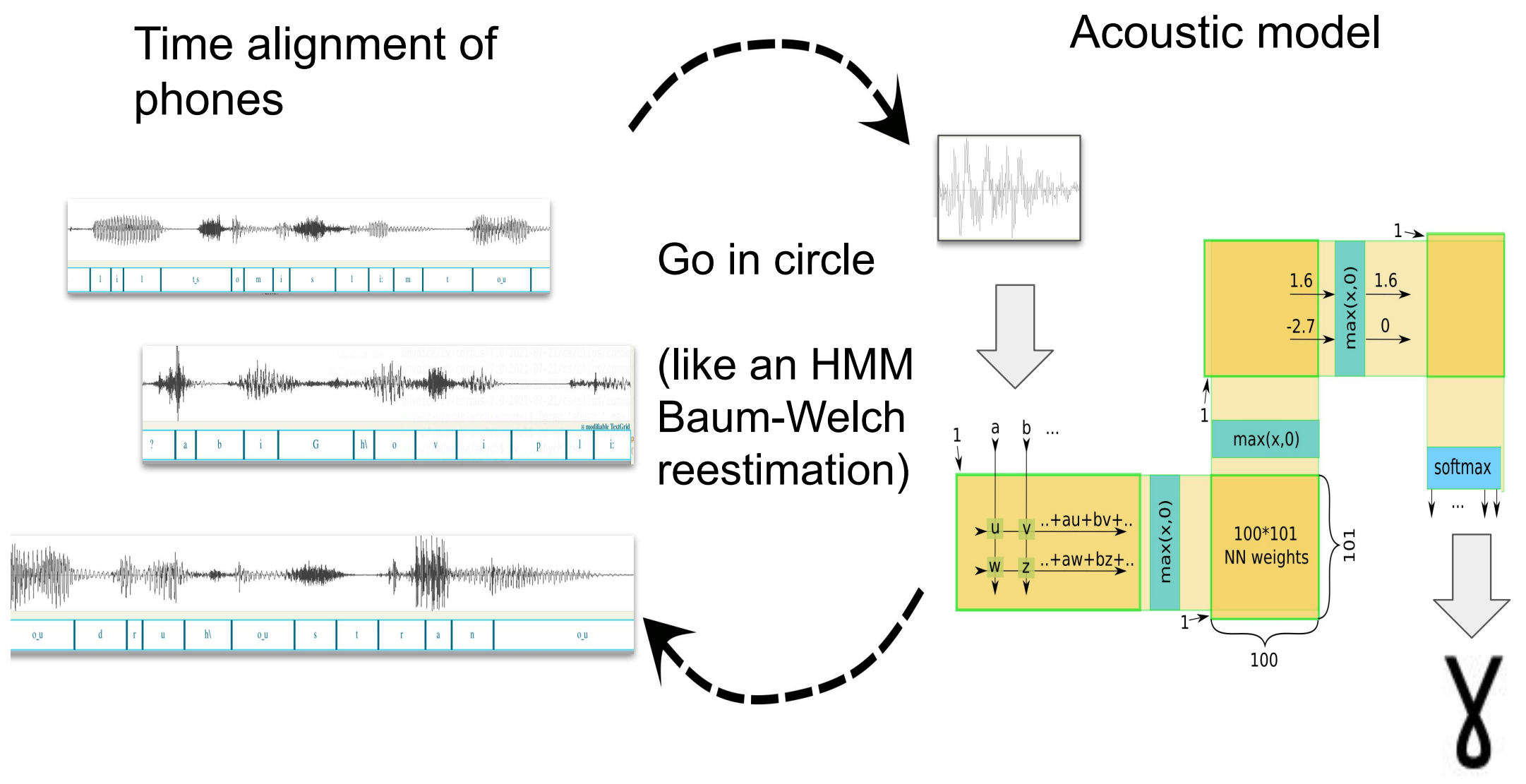
Software, documentation, ICPhS article, discussions forum, slides, references, this poster as pdf.
The README explains installation for all platforms.

Prak Design Goals

- an open-source tool free for any use – MIT license for code, trained on free audio data
- functioning on Mac, Linux and Windows
- easy to install – low dependencies, only reliable dependencies which are (hopefully) here to stay
- simple architecture, preferably building on techniques from phonetics students' curricula
- usable from both GUI and command line
- using explainable and modifiable logic (rather than a trained blackbox) where possible
- automatic pronunciation variant selection

DNN Acoustic Model Training

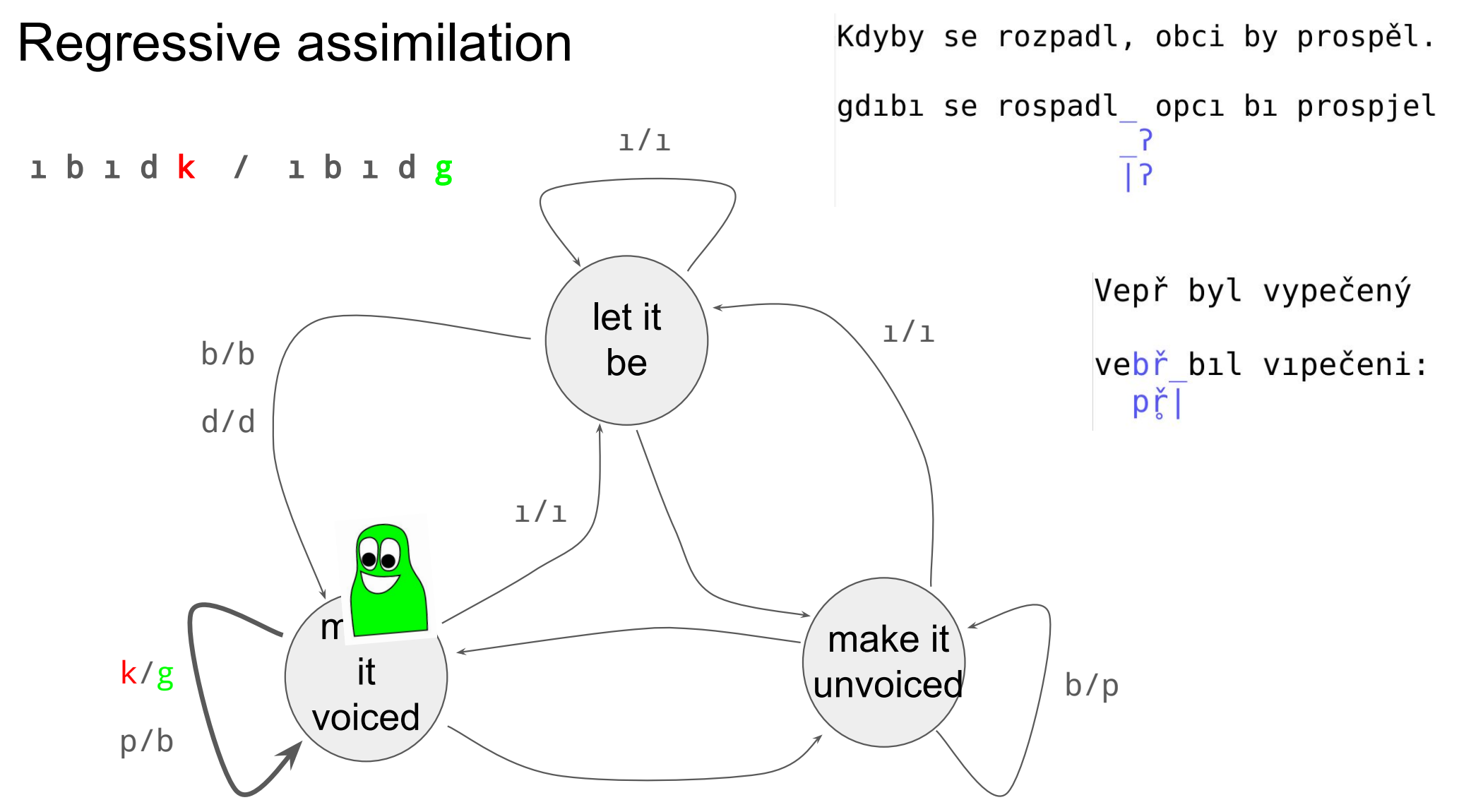
Alignment and model are gradually built in tandem by repeated training of the AM on the previously aligned data and re-aligning the same data by the new version of the model:



The initial alignment can be mostly wrong – the process can overcome this. We can thus train on CommonVoice which has only sentence-level alignment. Alternatively, part of the training data can have human-decided alignment which is kept constant during training.

Pronunciation Generator

String replacement rules cover foreign words. The regular part of Czech pronunciation is handled by a series of Finite State Transducers processing the phone string backward, taking care of assimilation or optional glottal stop insertions, for example:



```
Kdyby se rozpadl, obci by prospěl.

gdɪbɪ se rospadl_ opcɪ bɪ prospjel
                _ʔ
                |ʔ

Vepř byl vypečený

vebř_bɪl vɪpečeni:
  pr̥|
```

Both replacement rules and FSTs allow generation of multiple variants. These are resolved by the Acoustic Model match.

Precision

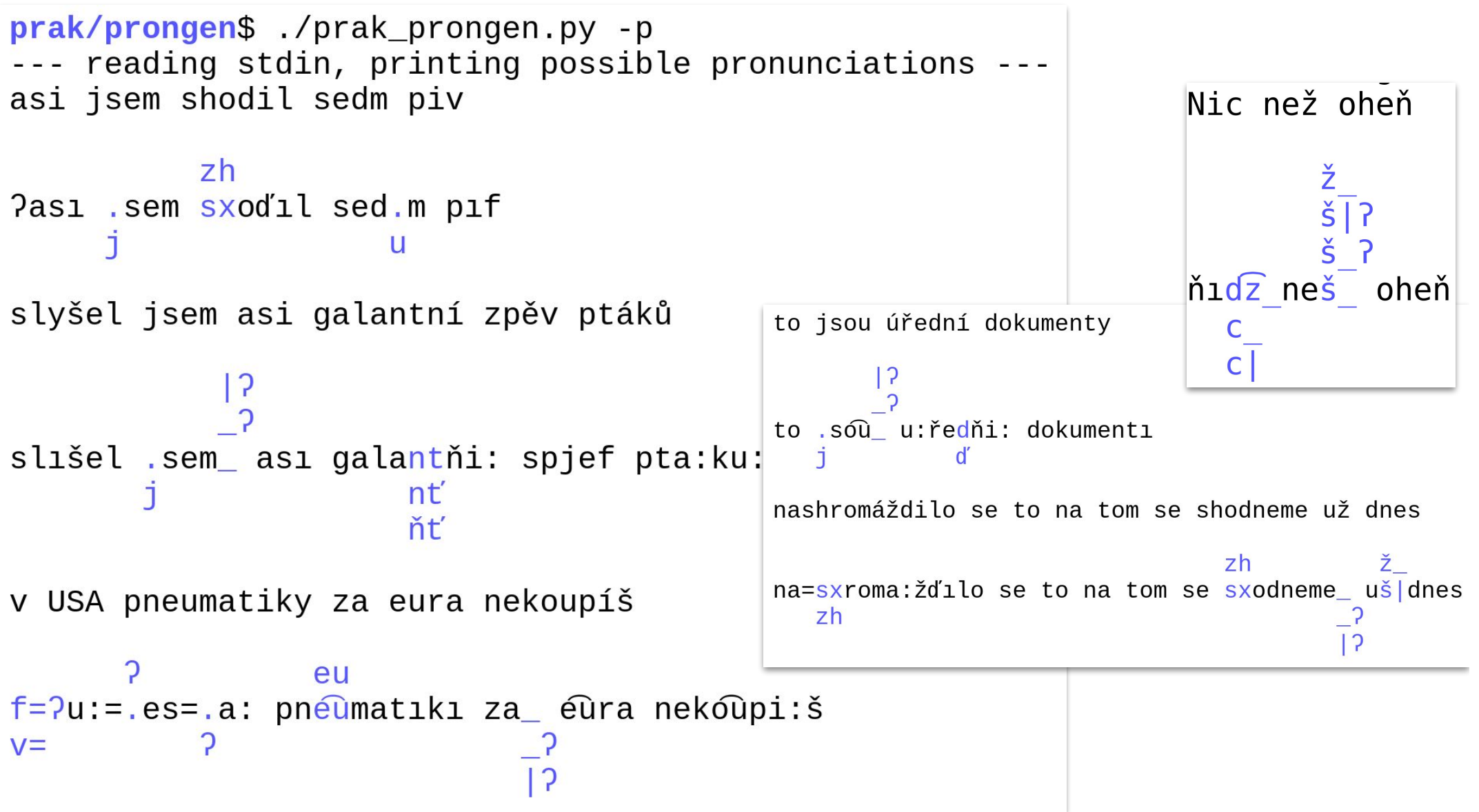
We measured the percentage of phone mismatch and boundary misplacement, comparing previously most used tool (Prague Labeler) to our Prak. The CV column is for Prak trained solely on the CommonVoice dataset. The FU column is for Prak trained on hand time-aligned private data and using HUBERT embeddings alongside MFCC (preliminary results, not covered in our article):

test	P.L. [1]	Prak CV	Prak FU
phone mismatch	6.61	1.88	1.12
match but misplace 0.1s+	**4.28**	**0.36**	0.04
match but misplace 0.2s+	3.22	0.09	0.00
mismatch or misplace 0.1s+	10.89	2.24	1.16

The CV column is for the general Prak. The FU column shows possible fine-tuning in a particular environment.

Command Line prongen Tool

For a given Czech text, the pronunciation generator creates a "sausage graph" suggesting variants at some places (here in blue, these are resolved by the AM):



```
prak/prongen$ ./prak_prongen.py -p
--- reading stdin, printing possible pronunciations ---
asi jsem shodil sedm piv

         zh
ʔasɪ .sem sxoďɪl sed.m pɪf
     j               u

slyšel jsem asi galantní zpěv ptáků

           |ʔ
           _ʔ
slɪšel .sem_ asɪ galantňi: spjef pta:ku:
       j              nť
                      ňť

v USA pneumatiky za eura nekoupíš

      ʔ         eu
f=ʔu:=.es=.a: pnéumatɪkɪ za_ éura nekóupi:š
v=         ʔ                _ʔ
                            |ʔ
```

```
to jsou úřední dokumenty

      |ʔ
      _ʔ
to .sóu_ u:ředňi: dokumentɪ
   j        ď

nashromáždilo se to na tom se shodneme už dnes

                                   zh        ž_
na=sxroma:žďɪlo se to na tom se sxodneme_ uš|dnes
   zh                                    _ʔ
                                         |ʔ
```

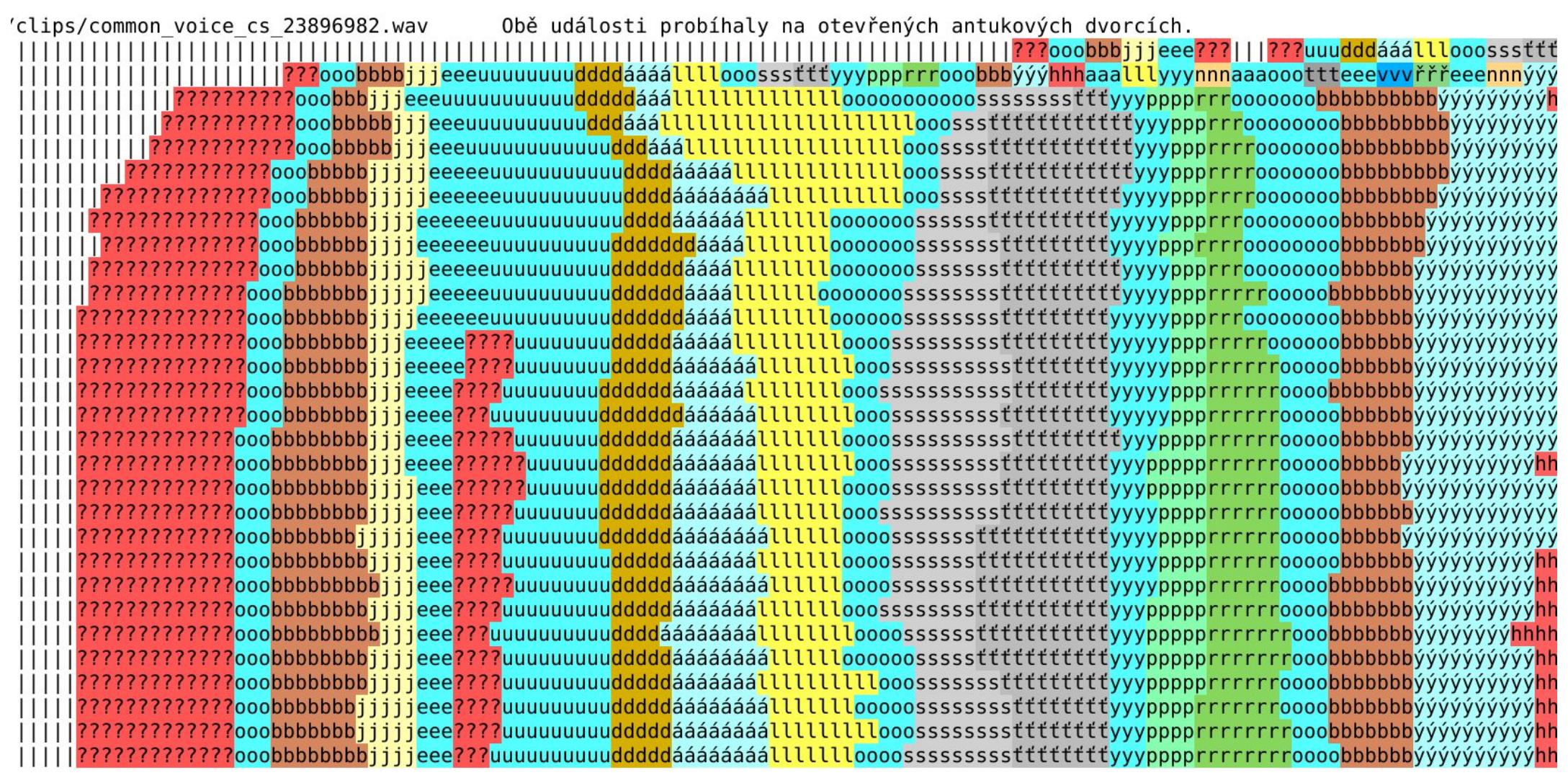
```
Nic než oheň

      ž_
      š|ʔ
      š_ʔ
ňɪdz_neš_ oheň
  c_
  c|
```

Alignment Emerging in Training

Each line is phone alignment in one train-align cycle:

"obě události probíhaly ..."



References

[1] P Pollák, J Volín, and R Skarnitzl. "HMM-based phonetic segmentation in Praat environment". In: *The XII International Conference Speech and Computer - SPECOM*. 2007, pp. 537–541.

This poster was presented at ICPhS Prague, 2023.

Future work: Try transformers as the AM. Try more languages. Try even tighter integration with Praat. The whole Prak was in fact designed as an easy to understand baseline. It has practical value as is but we hope that students of phonetics will build on it and try new ideas.

LaTeX TikZposter