AKADEMIE DER
WISSENSCHAFTEN
IN HAMBURG

Forced alignment using underspecified transcription for underresourced languages

Alexandre Arkhipov

Overview

1. Use cases for Audio Mining in the context of endangered languages/language documentation
2. (Mis-)using **WebMAUS** for phone-level segmentation in ‘untrained’ languages
3. Proposal for a general underspecified model based on larger sound classes

Audio Mining and endangered languages

A user perspective on Audio Mining in the context of:

- Linguistics
  - language documentation & description, linguistic typology, phonetics & phonology, multipurpose digital corpora
- Extracting acoustic features
  - to answer questions in phonetics & phonology

  E.g. is vowel length contrastive? is there voicing of stops in VCV position? what are the acoustic effects of feature X?..
- Working with audio data from endangered languages
  - limited data (several hours to max. several tens of hours)
  - legacy data, variable sound quality
  - no or few native speakers/consultants, few experts

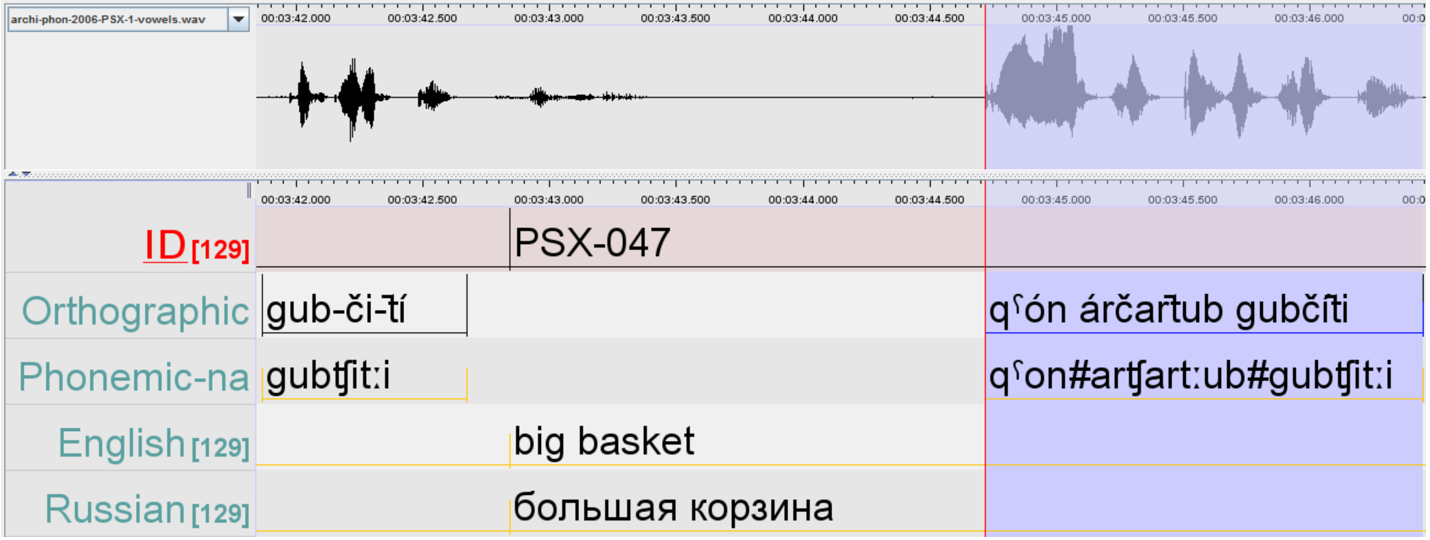
Audio Mining and endangered languages

Approaches to using phone-level forced alignment

- Input data:
    - several hours of audio per language
    - with broad (phonological) transcription and translation
    - time-aligned at phrase (sentence) level in ELAN or Praat
- Target output:
    - automated phone-level segmentation to reduce manual workload
    - (still with manual post-processing)

Audio Mining and endangered languages

AKADEMIE DER
WISSENSCHAFTEN
IN HAMBURG



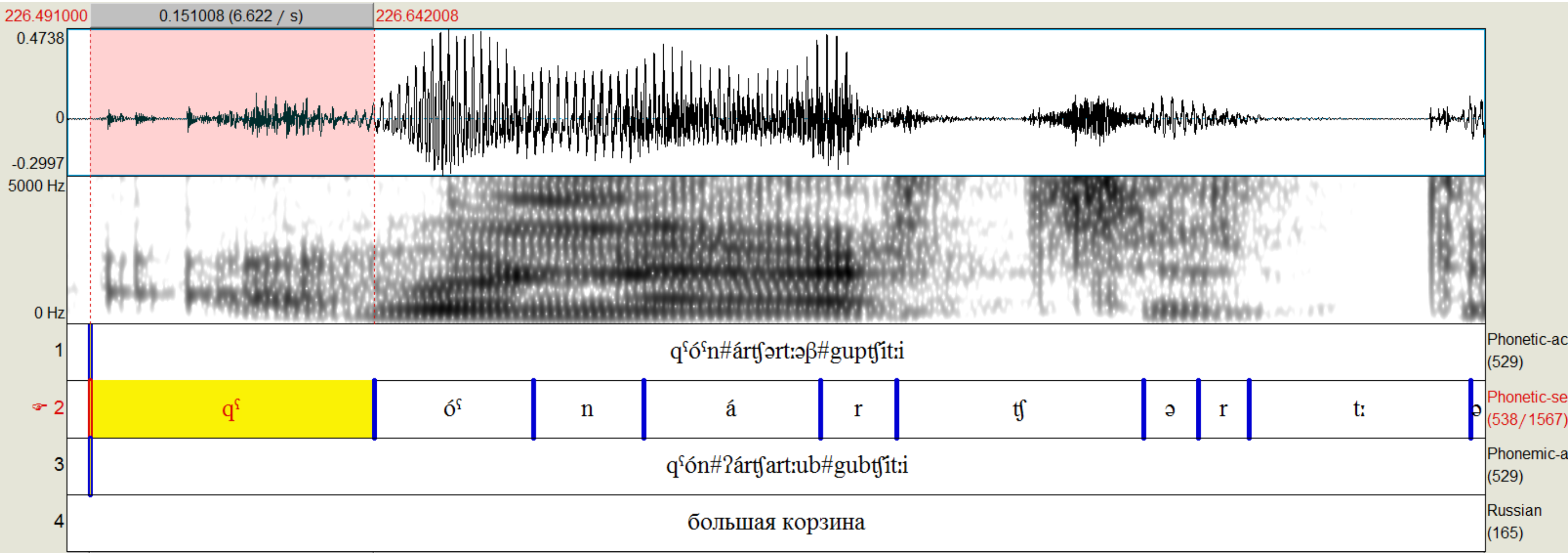
ID [129]: PSX-047

Orthographic: gub-či-ꞁ̄tí | qˤón árčarꞁ̄tub gubčíꞁ̄ti

Phonemic-na: gubʧiːi | qˤon#arʧartːub#gubʧitːi

English [129]: big basket

Russian [129]: большая корзина



1 qˤóˤn#árʧərtːəβ#guptʃitːi — Phonetic-ac (529)

2 qˤ | óˤ | n | á | r | ʧ | ə | r | tː | ə — Phonetic-se (538/1567)

3 qˤón#ʔárʧartːub#gubʧitːi — Phonemic-a (529)

4 большая корзина — Russian (165)

Audio Mining and endangered languages

Approaches to using phone-level forced alignment

- Input data:
  - several hours of audio per language
  - with broad (phonological) transcription and translation
  - time-aligned at utterance (sentence) level in ELAN or Praat
- Target output:
  - automated phone-level segmentation to reduce manual workload
  - (still with manual post-processing)
- **Problem**:
  - lack of trained acoustic models for specific languages
  - suitability of general (language-independent) models?

Using WebMAUS General

Here & below the **WebMAUS General** web-service was tested in action.
https://clarin.phonetik.uni-muenchen.de/BASWebServices/

1/ “Chunk preparation” tool (from ELAN source)

2/ Running WebMAUS General with the **language-independent (sampa) model**

- Quoting from the help page, “[t]he HMM set of language 'sampa' consists of the set union of all HMM sets of all languages known by MAUS. To cover all SAMPA symbols the missing symbols are then mapped to the phonetically nearest HMM”

Using WebMAUS General

Preliminary test: a very short fragment in Alutor
(a Chuckchee-Kamchatkan language, Russia)

- ❖**Result**: acceptable in general, not ideal for stops
- ❖Uvular /q/: left border often set too late (leaving out much of the closure)
- ❖**Reason?** Possibly rare or missing /q/ in the model
- ❖**Workaround**: Replace /q/ with another stop, e.g. /t/
- ❖**Result**: **Better accuracy** for left border
  (and generally not worse for the right border)

==> Hmm let's try again!

Working around lack of training data

- Dealing with ‘exotic’ and underresourced languages, we will always have sounds missing or underrepresented in existing trained models
- For the purposes of segmentation (not recognition!), some phonetic features are much more relevant than others, e.g.
  - ❖ major class (vowel, sonorant or obstruent)
  - ❖ manner of articulation, voicing >> place of articulation

Proposal:

To train a general model on **classes of sounds** acoustically similar w.r.t. transitions between classes

(cf. sequences like “voiceless stop-vowel”, “vowel-sonorant”, “voiced fricative-vowel”, etc.)

(Mis-)Using WebMAUS General

- **Idea:** while we (I) don’t have such a model, let’s (mis-)use the existing one
- **Approaching** a model with underspecified place of articulation for consonants:

Replace consonants of all places of articulation with coronals

(Why coronals? They are the best represented cross-linguistically)

- **Main test:** a fragment in Kamas (an extinct Samoyedic language, Russia)

Tape recordings of the last speaker, Klavdiya Plotnikova, made in 1964–1970, are now being transcribed in the INEL project

(Mis-)Using WebMAUS General

- **Data:** 195 seconds of SU0207 item (AEDKL archive)
  Recorded on tape in 1964, digitized in 2010 (48k/16bit)
- **Baseline transcription:** phonological, time-aligned in ELAN at utterance level
  - rendered into IPA
  - fed to Chunk preparation tool
  - KAN tier converted into SAMPA
  - fed to WebMAUS General to obtain a Praat TextGrid

  ⇒ **Baseline alignment**
- **Modified transcription:**

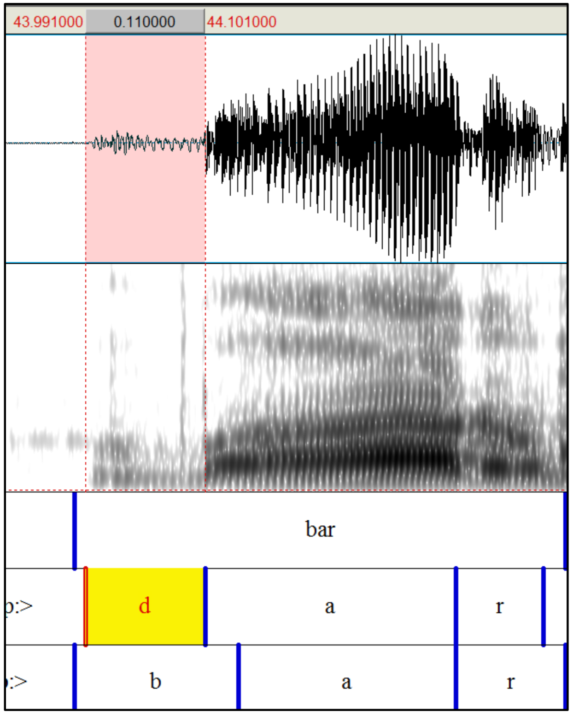
  ⇒ **Modified alignment**

b, g	> d	**b^j**	> d^j	**v**	> z
p, k	> t	**p^j**	> t^j	**ʃ**	> s
m, ŋ	> n				

Evaluating differences between baseline vs. modified alignment

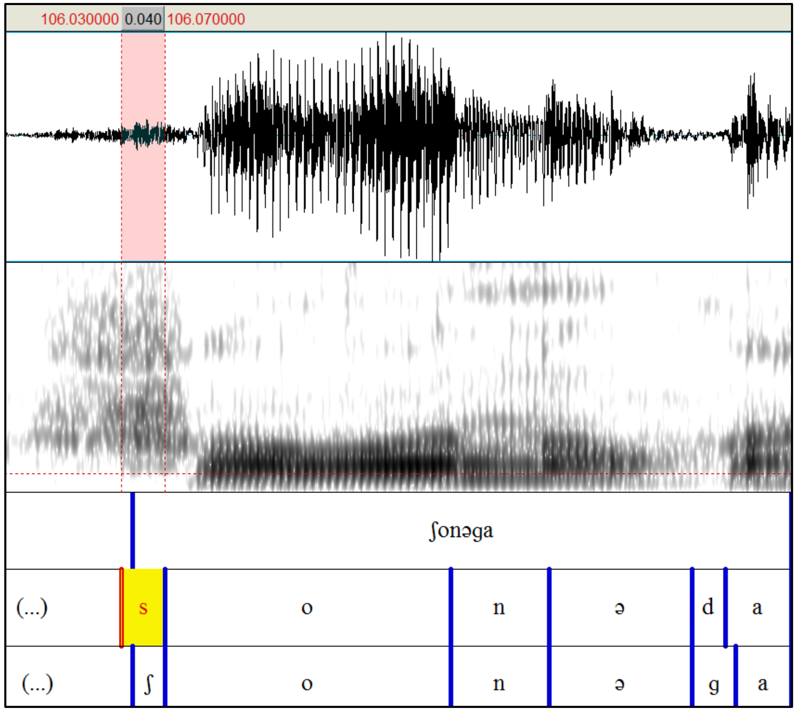
General remarks on results of forced alignment

- In most cases, phones are identified correctly
- Time precision is limited to 10 ms due to the nature of the algorithm
- When borders are far from optimal, a small improvement is not necessarily significant; however all will be classified as improvements.



[< left]
The alignment of /b/ (disguised as /d/) is close to optimal, given 10 ms precision, and significantly better than the baseline (lower tier)

[right >]
The modified alignment still captures only a portion of the total duration of the fricative



Evaluating differences between baseline vs. modified alignment

Sound \ Eval.	=	0	+−	−+	−	−−	+	++	Total
b > d	2	1		4	3	2	**33**		45
b^j> d^j		1			1				2
p > t	1	2					**7**		10
p^j > t^j							**3**		3
g > d	3	2	1	1	9		11		27
k > t	2	4			12	3	13	1	35
m > n	5	3		1	10	3	14		36
m (mn) > n (nn)				1	3		1		5
n (mn)		*(1)*			*(2)*		*(2)*		*(5)*
b (bd) > d (dd)					**3**				3
d (bd)					***(3)***				*(3)*
m (mb) > n (nd)	1	2			**8**				11
b (mb) > d (nd)	2	4		2	**2**		1		11
ŋ (ŋg) > n (nd)	1								1
g (ŋg) > d (nd)					**1**				1
Total	**17**	**19**	**1**	**9**	**52**	**8**	**83**	**1**	**190**

Evaluating differences between baseline vs. modified alignment

- **Surprise: In 84 of 190 instances, accuracy improved** after intentionally “corrupting” the transcription.

Unlike with Alutor, this can hardly be ascribed to rare or missing sounds, since most of the cases concern the fairly common /b, p, g, k, m/.

- **In labial stops /b, p/, 43 of 49** cases with unambiguous evaluation are classified as **improvement** (vs. 6 as deterioration).
- **In velar stops /g, k/ and nasal /m/**, the effect is seemingly **random**: resp. 25+ : 24– and 14+ : 13–.
- **Sequences with a second plosive (/bd, mb/)** clearly tend to show **deterioration:** 14– : 1+.

Labels for alignment change evaluation (see table above)

=	no change	**+**	improvement
0	change with no clear improvement/deterioration	**+ +**	improvement affecting non-adjacent phones
+ –	left border better/right border worse	**–**	deterioration
– +	left border worse/right border better	**– –**	deterioration affecting non-adjacent phones

Conclusion

- Results of the forced alignment may depend in unexpected ways on the transcription, including improved results from a “corrupted” transcription
- For the given use case (forced alignment with available chunked transcription), it’s worth trying to train a general model less sensitive to language-specific properties of phones — e.g. by using classes based on similarities of acoustic transitions
- (Or, alternatively, a post-processing module which could amend some of the shortcomings of the existing models)

INEL project

Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages

- Project aimed at creating corpora mainly from archival data for several selected languages of different families in Northern Eurasia
- 18 years (2016 – 2033)
- 3 years x 1 linguist per language/variety
- Principal Investigator: Prof. Dr. Beáta Wagner-Nagy
- Application by: Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Rießler, MA Timm Lehmberg, MA Hanna Hedeland
- Financing: the project is funded within the framework of the Academies' Programme, and coordinated by the Union of the German Academies of Sciences and Humanities
- Research: Institute for Finno-Ugric/Uralic Studies (IFUU) at Universität Hamburg
- Infrastructure: Hamburg Center for Language Corpora (HZSK)
- Homepage: https://inel.corpora.uni-hamburg.de/

References

- AEDKL
  Archive of Estonian Dialects and Kindred Languages (TÜEMSA), University of Tartu
  http://www.murre.ut.ee/arhiiv
- ELAN
  Wittenburg, P., Brugman, H., Russel, A., Klassman, A., Sloetjes, H. 2006.
  ELAN: a Professional Framework for Multimodality Research.
  http://tla.mpi.nl/tools/tla-tools/elan
- Praat
  Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.30, retrieved 22.07.2017 from http://www.praat.org/.
- WebMAUS
  Kisler, T., Reichel, U. D. and Schiel, F. 2017. Multilingual processing of speech via web services. // Computer Speech & Language, V. 45, September 2017, pp. 326–347.
  https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSGeneral
- Strunk, J., Schiel, F. and Seifart, F. 2014. Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS. // LREC 2014. Reykjavik. Pp. 3940–3947.