Autophon.org
An online tool for automatic phonetic annotation

Nathan Joel Young[1], *Kaosi Anikwe*[2]
[1]*University of Oslo,* [2]*University of Nigeria Nsukka*
n8.young@gmail.com, anikwehenryasa@gmail.com

Abstract

Autophon (autophon.org) is a free web application that facilitates automatic phonetic annotation. It supports both Nordic and low-resource languages, addressing critical gaps in phonetic research. The platform reduces technical barriers, saving time for students and researchers. Rigorous validation ensures the accuracy and reliability of its tools, making it a valuable resource for the linguistic community.

Introduction

We introduce Autophon (autophon.org), a free worldwide web application designed to facilitate phonetics research through forced alignment technology. This initiative simplifies phonetic annotation by using neural networks via systems like The Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017). It automatically timestamps speech recordings, aligning them with transcriptions to produce phonetic annotations in Praat (Boersma & Weenink, 2024), optimizing user inputs and expediting tasks that traditionally required significant manual investments. This is especially beneficial in regions like Scandinavia where the high labor cost for manual phonetic annotation can eat up research budgets.

The project's mission is to democratize forced alignment tools, evident in its free accessibility and focus on Nordic languages. In bypassing complex command-line programs, it empowers students and researchers to use speech analysis tools efficiently. Security and privacy of user data are ensured, with files deleted after alignment. Despite being in *beta*, efforts towards stability and bug fixes are ongoing.

Background

Forced alignment technology automates the alignment of speech recordings with their transcripts, ensuring precise synchronization of phonetic segments. Traditionally, this was done manually, which was labor-intensive and prone to errors. We offer a short review of this technology here; however, for a more exhaustive review, refer to Young & McGarrah (2023). The early 2000s saw the rise of Gaussian Mixture Models in computational linguistics (Young, Woodland, & Byrne, 1993), which phoneticians used to construct forced aligners such as FAVE-Align (Rosenfelder, Fruehwald, Evanini, & Yuan, 2011) and The Multilingual Annotation and Alignment Tool MAUS (Schiel, Draxler, & Harrington, 2011).

Recently, the field has shifted over to Deep Neural Networks on frameworks like Kaldi (Povey et al., 2011), leading to the Montreal Forced Aligner (MFA). Autophon, the web app discussed here, currently uses MFA vers. 1.0 for its backend operations.

WebMAUS marked a different milestone by offering this technology on an online portal (Schiel, Draxler, & Harrington, 2011). While certainly enhancing accessibility for researchers, it has also faced challenges like codec specificity, compatibility errors, and limited technical support. Other online portals providing access to forced alignment

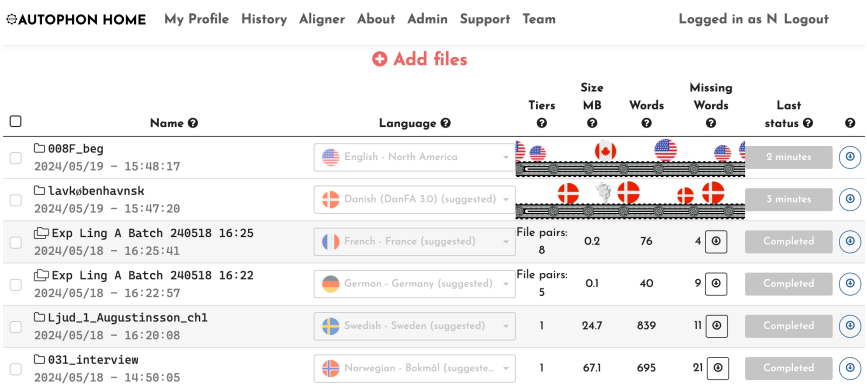

Figure 1. A screenshot of the Autophon application.

tools include the University of Wisconsin-Milwaukee Forced Aligner[1] and the North Carolina State University Forced Alignment Service[2]. While valuable, these platforms often suffer from the same weaknesses as WebMAUS.

Autophon distinguishes itself by offering an intuitive UX and a codec-agnostic engine, improving accessibility, efficiency, and reliability. In the ensuing prose, we highlight the app's three main focuses: (1) integrating diverse forced alignment tools into a user-friendly interface, (2) developing new models for Nordic languages, and (3) bootstrapping models for low-resource languages.

Focus 1: User ease

Access first

Figure 1 illustrates a screenshot of Autophon's user interface (UX). By *access first* we mean that even the most accurate phonetic tools are of little utility if (a) the barrier to access is too tedious or time-consuming and/or (b) validating accuracy is too difficult. The abundance of command-line tools in the field necessitates a need to remove barriers, facilitate user access (UX), and offer support. What is more, researchers often face a *paradox of choice* where too many options can lead to a sort of decision paralysis that hinders productivity (Iyengar & Lepper, 2000; Schwartz, 2004). Evaluating and testing multiple programs becomes a task unto itself, diverting time and effort from primary research objectives.

We try to meet this challenge with Autophon by consolidating various tools into a unified platform. It currently features MFA vers. 1.0 with plans to incorporate other tools like MFA vers. 3.0, FAVE/Penn Forced Aligner, LG-FAVE, and – further in the future – Wav2vec. By providing diverse tools in one place, Autophon streamlines the selection process, allowing researchers to efficiently identify and utilize the most effective tools without being overwhelmed. This empowers users to focus on research rather than the complexities of tool implementation and evaluation.

Search engine optimization

Searching "online forced aligner" or "automatic phonetic annotation Nordic languages" on Google brings up Autophon as a top result. Such effective search engine optimization (SEO) facilitates accessibility and visibility, ensuring that researchers – regardless of network connections – can easily find and use the app. In a place like Scandinavia where research networks often are siloed and invitation-only, SEO makes access

[1] https://web.uwm.edu/forced-aligner

[2] https://phon.chass.ncsu.edu/cgi-bin/step7.cgi

more horizontal, fostering a broader community of scholars.

Intuitive UX

Autophon's intuitive user interface (UX) helps reduce the learning curve and time investment for users. It is built to streamline the onboarding process, enabling users to quickly understand and use the application's features. This is particularly important in phonetics, where technical barriers often deter students, especially undergraduates. But even seasoned researchers require efficient tools and workflows as they face off against performance pressure in academia. Phonetics as a field requires exorbitant time for data collection, annotation, and processing. In the neoliberal academic environment and an increasingly market-driven research funding milieu, this reduces the "competitiveness" of phonetics compared to less technical humanities disciplines. The UX is designed to minimize the time spent on processing, allowing phoneticians to reallocate their time to empiric inquiry and theory.

Tech support

Autophon provides tech support, setting it apart from traditional open-source web and desktop programs. Desktop applications require individualized troubleshooting, which often makes them resource-intensive and subject to unpredictable funding streams. In contrast, Autophon's web-based platform allows for scalable problem-solving and what we call *altruistic debugging*: when one user's problem is fixed, it benefits all users. The approach makes the app resilient to various operating systems and their technical issues. Further, Autophon's commitment to tech support enhances the user experience and contributes to a more inviting environment for phonetics research.

Accuracy metrics

Significant resources have been dedicated to validating models against manual corrections, with current validation statistics available for Danish, English, Norwegian, Swedish. Validation helps researchers make and justify their choices, ensuring accountability and replicability in their work. Standard established metrics for validation are used such as median and mean offset time and percentage of offsets that fall within ten and 20 milliseconds from the manual gold standard (Young & McGarrah, 2023, p. 113).

Focus 2: Nordic languages

Although the mainland Scandinavian languages are considered high-resource by most measurements, forced alignment technology has lagged for them, being mostly only available for "super-resource" languages like English and French (c.f. Lindh, 2007a). It is our view that this may be one reason behind the scarcity of large-scale phonetically annotated corpora of spontaneous Scandinavian. This section briefly examines the current state of forced alignment for Nordic languages, our view on the implications for linguistic research, and our efforts to address these challenges.

Current state in Scandinavia

Advancements in automatic phonetic annotation for English have coincided with numerous public corpora of phonetically annotated *naturally occurring spontaneous speech* (Kendall & Farrington, 2023; Pitt, Dilley, Johnson, Kiesling, Raymond, Hume, & Fosler-Lussier, 2007; Coleman, Baghai-Ravary, Pybus, & Grau, 2012). In contrast, Nordic languages have remained underserved. For example, the Swedish Dialect Database *SweDia 2000* only has its read-aloud section phonetically annotated (Eriksson, 2004, p. 39). The RUNDKAST corpus, a Norwegian Broadcast News Speech Corpus, includes only 50 minutes of annotated read-aloud speech and ten minutes of spontaneous speech (Amdal et al., 2008). The DanPASS corpus for Danish

provides phonetic annotation only at the syllabic level for spontaneous speech (Grønnum, 2009). To our knowledge, the only publicly available corpus of a Nordic language with phonetically annotated spontaneous speech is the ten minutes from RUNDKAST, highlighting a significant gap in resources.

With Autophon, the aim is to help close this gap by lowering the technological barrier to annotating naturally occurring Scandinavian speech. This is particularly crucial for two fields; namely, (1) research on language variation and change and (2) forensic speaker comparison. We offer a case in point below for each.

Pertaining to variation and change, recent public inquiries were made into Swedish multiethnolect and its phonetic spread into standard Swedish (SVT Nyheter, 2024; Sveriges Radio, 2024). In response to this inquiry, Swedish academia had few studies to refer the public to (cf. Gross, Boyd, Leinonen & Walker, 2016; Gross, 2018; Young, 2021). Pertaining to forensic speaker comparison (Jessen, 2008), reference corpora of urban Swedish are absent, which increases the risk of false positives because typicality must be calculated on more modest regional data like, e.g., Lindh (2007b). The current proliferation of telephone fraud cases in Sweden (SVT Play, 2024) underscores the acute need for phonetically annotated reference corpora that include spontaneous speech from speakers of newer urban sociolects.

Autophon's contributions

Autophon is one small cog within the research enterprise, merely a tool. However, this tool provides a unified space where researchers of Nordic phonetics and phonology can access the latest forced alignment tools, significantly reducing the time and effort required to navigate and evaluate multiple programs. This inclusive approach not only simplifies the research process but also empowers researchers to focus on their primary objectives rather than the complexities of tool vetting and implementation.

Focus 3: Smaller languages

With this tool, we seek to bridge the gap in forced alignment technology for low-resource languages by offering *bootstrap models*, a method that has proven effective in previous studies (Coto-Solano & Solórzano, 2017; Young & McGarrah, 2023). Bootstrapping means adapting existing models from one language to a second genealogically close language.

Methodology

The methodology for bootstrapping involves several steps. Initially, existing models that are trained on a resource-rich language are adapted to a new target language. This adaptation is done by leveraging phonetic and linguistic similarities between the source and target phonemes. Subsequent alignments produced by the bootstrap model can then be manually corrected and used to train native models for the target language.

Young (2017a) demonstrated the bootstrap approach with a Swedish-language adaptation of FAVE-Align, showing that it is robust and reliable for both spontaneous and read-aloud Stockholm Swedish. This same procedure was later also conducted for Danish (Young, 2017b). By starting with the existing Hidden Markov Models trained on English in FAVE-Align (Rosenfelder et al., 2011), significant reductions in manual segmentation time were achieved with a substantial proportion of boundaries falling within acceptable error margins compared to manual alignments (Young & McGarrah, 2023).

Faroese and Icelandic

Autophon has adopted this bootstrapping approach to develop aligners for Faroese and Icelandic based on a Norwegian language model (Young, 2020). These aligners offer a starting point for

researchers working on these languages, enabling them to conduct initial phonetic annotations, which can later be refined and corrected. This iterative process will eventually help develop fully native models for these languages.

Limitations and future directions

No formal validation statistics are available for our bootstrap models, so we are seeking collaborations with researchers to assist in calculating them. By working together, the goal is to gather enough manually corrected data to construct robust models that meet established accuracy benchmarks. As part of the initiative to bootstrap models for low-resource languages, prototypes for Elfdalian and Greenlandic are also being developed by leveraging existing resources (e.g., Oqaasileriffik, 2007; Steensland, 2021).

Conclusion

The Autophon project is a forward-thinking approach to overcoming challenges in phonetic research, particularly for Nordic and other low-resource languages. The platform prioritizes access and use, reducing the learning curve for students and researchers. By integrating multiple forced alignment tools into a user-friendly web application, it minimizes complexities typically associated with phonetic annotation, allowing users to focus on their core research objectives.

A significant contribution of this project is its support for educational environments. The intuitive user interface and comprehensive resources make it ideal for teaching phonetics and linguistics, enabling students to engage with advanced speech analysis tools without technical barriers. This accessibility enhances learning and encourages deeper exploration of phonetic phenomena.

For researchers, the project offers a streamlined phonetic annotation process, saving substantial time, effort, and money. The focus on creating models tailored to Nordic languages addresses a critical gap, with bootstrap models facilitating initial alignments and further refinement. This iterative process fosters collaboration and continuous improvement, contributing to high-quality linguistic resources.

The commitment to validation underscores the project's dedication to supporting the research community. By validating models against manual corrections and adhering to established metrics, the platform ensures reliability and trustworthiness. Ongoing validation efforts demonstrate a strong commitment to quality.

In summary, this initiative represents a significant advancement in phonetic research tools, offering substantial benefits for both teaching and research. By bridging the gap between computational methods and the practicalities of access, Autophon supports research efficiencies, the aim of which is to foster a more relevant and rigorous field of phonetics.

Abbreviations

MFA – Montreal Forced Aligner
SEO – search engine optimization
UX – user interface

Acknowledgements

Autophon was funded in part by a grant from the Swedish Academy, a grant from the Department of Linguistics and Scandinavian Studies at The University of Oslo, and the European Union's Horizon 2020 Marie Sklodowska-Curie grant agreement No 892963.

References

Amdal, I., Strand, O. M., Almberg, J., & Svendsen, T. (2008). RUNDKAST: an Annotated Norwegian Broadcast News Speech Corpus. *Proc. of LREC*.

Boersma, P., & Weenink, D. (2024). *Praat: doing phonetics by computer* (6.4.12). http://www.praat.org/

Coleman, J., Baghai-Ravary, L., Pybus, J., & Grau, S. (2012). *Audio BNC: The*

audio edition of the Spoken British National Corpus. University of Oxford.

Coto-Solano, R., & Solórzano, F. (2017). Building and evaluating a forced aligner for an endangered language: Bribri. *Proc. of the 2nd Workshop on the Use of Comp. Methods in the Study of Endangered Languages* (pp. 1–9).

Eriksson, A. (2004). SweDia 2000: A Swedish dialect database. In P. J. Henrichsen (Ed.), Babylonian confusion resolved. *Proc. of the Nordic symposium on the comparison of languages* (pp. 33–48).

Grønnum, N. (2009). DanPASS – A Danish phonetically annotated spontaneous speech corpus. *Speech Communication*, *51*, 594–603.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995–1006.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass, 2*(4), 671–711.

Kendall, T., & Farrington, C. (2023). *The Corpus of Regional African American Language. Ver. 2023.06*. Eugene, OR: The Online Resources for African American Language Project.

Lindh, J. (2007a). Semi-automatic aligning of Swedish forensic phonetic phone speech in Praat using Viterbi recognition and HMM [unpub.manus.].

Lindh, J. (2007b). Preliminary Descriptive F0-Statistics for Young Male Speakers. *Working Papers, 52*, Lund University, Centre for Lang. & Lit, 89–92.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proc. of Interspeech*, 498–502.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech* (2nd release). Columbus, OH: Dept. of Psychology, Ohio State University.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Veselý, K. (2011). The Kaldi speech recognition toolkit. *Proc. of IEEE Workshop on ASR and Understanding*.

Oqaasileriffik. (2007). *IPA-Korpus-Sanik-TAAL: A Greenlandic Language Corpus with IPA Transcriptions*. https://github.com/Oqaasileriffik/ipa-ks

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). *FAVE Suite*.

Schiel, F., Draxler, C., & Harrington, J. (2011). Phonemic segmentation and labelling using the MAUS technique. In *Workshop New Tools and Methods for Very-Large-Scale Phonetics Research*, Philadelphia, PA.

Schwartz, B. (2004). *The Paradox of Choice: Why More Is Less*. Harper Perennial.

Steensland, L. (2021). Älvdalsk ordbok. In M. Wiss (Ed.), *Ulum Dalska*. https://dalsk.ordbok.gratis/

SVT Nyheter. (2024, April 2). *Språkforskaren: Staccato-rytmen sprider sig i svenska språket*. https://www.svt.se/kultur/sprakforskaren-staccato-rytmen-sprider-sig-i-svenska-spraket

SVT Play. (2024, February 7). *Uppdrag granskning: Bedragarna.* https://www.svtplay.se/video/KVk3QWv/uppdrag-granskning/bedragarna

Sveriges Radio. (2024, April 16). *Danska kungabesöket, OS-mode, Platons sista tid i livet.* https://sverigesradio.se/avsnitt/danska- kungabesoket-os-mode-platons-sista-tid-i-livet

Young, N. J. (2017a). *SweFA 1.0 – Forced Alignment of Swedish, Ver. 1.0*. https://github.com/ mcgarrah/LG-FAVE

Young, N. J. (2017b). *DanFA 1.0 – Forced Alignment of Danish, Ver. 1.0*. https://github.com/mcgarrah/LG-FAVE

Young, N. J. (2020). *NoFA 1.0 – norsk modell for forced alignment, Ver. 1.0*. https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-59/

Young, N. J., & McGarrah, M. (2023). Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics, 46*(1), 105–131.

Young, S. J., Woodland, P. C., & Byrne, W. J. (1993). *HTK: Hidden Markov Model Toolkit, Ver. 1.5*. Cambridge: Cambridge University.