A quantitative analysis of human corrections to forced alignment and their practical consequences

Nil Akintürk[1], Cesko Voeten[12]
[1]University of Amsterdam, Amsterdam Center for Language and Communication, [2]Fryske Akademy

A common approach to segmenting speech into phones is to use *supervised forced alignment*, where recordings are manually annotated at the sentence or word level, and a computer tool takes these annotations (in IPA, usually through G2P conversion) and aligns the phone boundaries to the speech signal. Both conventional wisdom and hard data hold that automated forced alignment should be hand-corrected post-hoc, due to inevitable errors. However, manual corrections are highly laborious; what's more, they are often done by multiple correctors (e.g. different student assistants), introducing additional variation.

The present paper is focused on this second source of variation. While it is well-established that a dataset with hand-corrections is superior to an uncorrected ('baseline') dataset (e.g. MacKenzie & Turton 2020), we extend this by studying variation *among* correctors. In addition, we assess the *practical consequences* for research in the domain of phonetic variation in change. Evanini, Isard, & Liberman (2009) have shown for *automatic formant extraction* that if a corpus is sufficiently large, formant-tracking errors are negligible in practice; we examine whether this might hold similarly for our research object of phone segmentation. Here, too, we explicitly take into account the additional factor of inter-corrector variation.

Our dataset consists of a total of 112,058 vowel tokens, of which 58,714 had been corrected by a total of six student assistants (all Linguistics majors with ample background in phonetics) plus their supervisor (an assistant professor of sociophonetics/sociophonology). These (non-)corrections were obtained from the Boarnsterhim Corpus (Kingma et al submitted), a recently-completed dataset containing sociolinguistic-interview and word-list data from 112 speakers from the Frisian municipality of Boarnsterhim. The corpus had been manually transcribed orthographically and aligned at the sentence level, and subsequently been aligned at the word and, in a second pass, phone levels using the Montréal Forced Aligner (McAuliffe et al 2017; note that this aligner is currently the top state of the art according to the comparison by Rousso et al 2024).

Our analysis consisted of both exploratory analysis and confirmatory hypothesis testing. Exploratorily, we investigated (i) the size of the disagreement between the aligner's automated boundaries and human corrections stratified by manner of articulation, (ii) the extent to which correctors disagree with one another on the same boundary type stratified by manner of articulation, and (iii) whether formant transitions at these boundaries can explain these disagreements. Confirmatory hypotheses tests assessed (iv) the extent to which corrections of the starting and end points of the same tokens significantly differed among correctors (using LMMs); (v) whether specific formant transitions predicted these corrections; (vi) whether speaker characteristics (specifically, sex and speaker identity) and vowel characteristics (position, consonant voicing, consonant PoA, consonant manner, vowel height, vowel backness) predicted these corrections; (vii) whether the corrections have *practical consequences* for the material conclusions that would be drawn by a sociophonetic analysis of these data's F1, F2, and F3 values (using GAMMs [Wood 2017], cf. Voeten, Heeringa, & Van de Velde 2022).

Our results show that both aligner-corrector and inter-corrector disagreement are largest for sonorants and vowels that are bordering silence, and the smallest where the vowel meets an obstruent, suggesting that there is an overlap between the cues that are used by the model and the humans annotators. Significant differences among human correctors were found; they are not significantly associated with formant transitions, but various effects of speaker and vowel characteristics are significantly associated with these inter-corrector differences. Perhaps most importantly, there were very little differences in GAMM fits between uncorrected data, corrected data overall, and corrected data split out by the different correctors. All of this suggests that while human corrections may be superior to unsupervised forced alignment, the high cost of human corrections may not justify the modest gains.

References

Evanini, K., Isard, S., & Liberman, M. Y. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. In *Proceedings of Interspeech 2009* (pp. 1655-1658).

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, *6*(s1), 20180061.

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

Kingma, M., Pinget, A-F., Heeringa, W. & Van de Velde, H. (submitted). The Boarnsterhim Corpus: A Frisian-Dutch bilingual speech corpus in apparent- and real-time. *Language Resources and Evaluation*.

Rousso, R., Cohen, E., Keshet, J., & Chodroff, E. (2024). Tradition or innovation: A comparison of modern ASR methods for forced alignment. *arXiv preprint arXiv:2406.19363*.

Voeten, C. C., Heeringa, W., & Van de Velde, H. (2022). Normalization of nonlinearly time-dynamic vowels. *The Journal of the Acoustical Society of America*, *152*(5), 2692-2710.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd edn.). Chapman & Hall/CRC.