

Literature Review on Forced Alignment Challenges in Diverse Languages

Shaima Dayili

George Mason University
Sdayili@gmu.edu

Abstract

Forced alignment (FA) is a foundational technique in speech science that enables the automatic temporal alignment of transcriptions with speech signals. Although FA systems perform reliably for high-resource languages, their accuracy degrades when applied to typologically diverse and under-resourced languages. This literature review suggests that the current constraints of forced alignment are not merely technical issues but are rooted in fundamental representational issues. By synthesizing research from multilingual processing, language documentation, and articulatory phonology, this review illustrates how traditional FA systems often mask critical cross-linguistic differences in articulation and temporal coordination. Through an analysis of empirical studies involving WebMAUS and cross-language alignment alongside theoretical work on gestural models, this review argues for the integration of gestural scores into computational frameworks. Such an integration provides a principled solution to cross-linguistic variability by capturing the physical realities of speech production that symbolic and purely acoustic models miss.

1. Introduction: Forced Alignment and Linguistic diversity

Forced alignment (FA) refers to the automatic determination of temporal boundaries for linguistic units, such as phones or words, within a speech signal based on a written transcription. FA is widely used in corpus phonetics, sociophonetics, and language documentation, where manual annotation is often impractical. Most widely utilized FA systems are built on acoustic models trained on high-resource languages, particularly English, and assume relatively stable mappings between symbolic phonological units and their acoustic realizations.

A growing body of research demonstrates that these assumptions do not generalize well to typologically diverse or under-resourced languages. Research evaluating the application of cross-language forced alignment indicates that while the alignment process is generally viable, its accuracy varies systematically across segment types and prosodic contexts [1], [2]. These findings imply that alignment errors are not merely random; rather, they reflect fundamental mismatches between the abstract speech representations utilized by FA systems and the actual physical production of speech across languages.

Standard FA systems treat phonological representations as sequences of discrete segments and implicitly assume that identical IPA symbols correspond to comparable articulatory events across languages. However, research in laboratory phonology demonstrates that the same symbolic category may correspond to distinct articulatory configurations and timing

patterns across languages and dialects [3], [4]. This representational gap motivates the central argument of this review: forced alignment accuracy is constrained by segment-based representations, and articulatory–gestural models provide a more appropriate theoretical foundation for cross-linguistic alignment.

2. Multilingual Forced Alignment Systems and Their Representational Limits

2.1. Web-Based Multilingual Alignment Frameworks

One major response to cross-linguistic variability has been the development of multilingual forced alignment services. The BAS CLARIN web services provide a modular infrastructure for multilingual speech processing, including grapheme-to-phoneme conversion and forced alignment [5]. These services make FA accessible to researchers working on languages for which no dedicated acoustic models exist.

Within this framework, WebMAUS has emerged as a widely used aligner in language documentation and corpus-based research. WebMAUS is designed to perform untrained forced alignment by combining existing acoustic models with language-specific grapheme-to-phoneme mappings [6]. Evaluations show that this approach can yield usable word- and phone-level alignments across typologically diverse languages, particularly when manual pre-alignment of larger units is available.

Despite these successes, alignment quality varies considerably across languages. [2] demonstrates that when English-based models are applied to conversational Kriol, alignment errors cluster around segments whose phonetic realization diverges from English norms. These findings indicate that multilingual FA systems remain constrained by the segment-based assumptions in their acoustic models.

2.2. Language-Specific Adaptations and Romanization Systems

To enhance the accuracy of alignment, researchers often implement symbolic adaptations tailored to languages. A notable example is the development of a specialized romanization system for Arabic varieties, specifically designed for integration with the WebMAUS aligner [7]. This system improves grapheme-to-phoneme correspondence and overall alignment accuracy by explicitly encoding phonological contrasts, such as vowel length and consonant gemination, the system improves.

While such adaptations highlight the necessity of language-specific expertise, they simultaneously reveal the inherent limitations of purely symbolic solutions. Even when using a refined romanization system, a single segment label remain a static representation that may fail to capture the diverse

articulatory timing or gestural coordination patterns found across different dialects.

3. Forced Alignment in Low-Resource and Cross-Linguistic Contexts

3.1. Cross-Language Alignment as a Practical Necessity

For many under-resourced languages, forced alignment necessarily relies on cross-language acoustic models. [1] shows that cross-language FA can substantially reduce annotation time in community-based linguistic documentation. Similarly, [2] finds that word-level alignment is often reliable for Kriol, while phone-level accuracy remains variable.

Based on the analyses in [1], [2], Figure 1 summarizes error distributions across segment types in cross-language forced alignment.

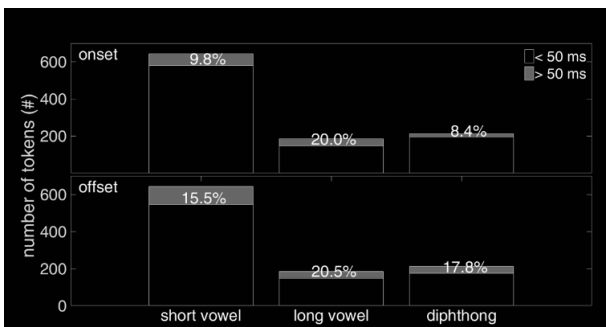


Figure 1. Onset and offset alignment accuracy by vowel type.

These studies demonstrate that alignment errors correlate with phonetic and prosodic properties of the target language, suggesting that differences in gestural organization underlie many alignment failures.

3.2. Semi-Automatic Strategies and Expert Intervention

To address the limitations of fully automated systems, researchers often implement semi-automatic alignment workflows that incorporate human linguistic expertise. These hybrid strategies often involve manually partitioning lengthy recordings into manageable segments to simplify the task for the aligner. Additionally experts may intervene to correct micro-pauses, where recognizers mistakenly start a word later than a human would, and perform iterative refinements on the output to ensure the alignment matches the actual phonetic realization of the target language [8]. Semi-automatic alignment for under-resourced languages demonstrates that human knowledge of phonetic structure can compensate for some system limitations.

However, these interventions operate around the representational problem rather than revolving it. They implicitly acknowledge that accurate alignment depends on understanding how speech is physically produced, even if that understanding is not explicitly encoded in the FA system itself.

4. Articulatory Phonology and the Gestural Representation of Speech

4.1. Theoretical Foundations of Articulatory Phonology

Articulatory Phonology provides a robust theoretical framework to overcome the representational constraints found

in traditional forced alignment systems. Instead of viewing speech as a linear string of discrete, static segments, this model conceptualizes speech as coordinated patterns of overlapping articulatory gestures. These gestures are defined as dynamic actions of the vocal tract characterized by spatial and temporal properties [9].

Within this framework, phonological contrasts are understood as emerging from differences in gestural coordination and constellations rather than from static symbolic units. This perspective is especially valuable for cross-linguistic analysis; it recognizes that while two languages may share a common phonetic label or IPA symbol, they may possess significant underlying differences in gestural timing and magnitude. By utilizing a gestural score, researchers can visualize these distinct realizations, providing a principled link between phonological intent and physical speech production that purely acoustic models often fail to capture.

4.2. The Gestural Score as a Cross-Linguistic Diagnostic Tool

The gestural score provides a visual and analytical representation of articulatory activity over time, displaying how gestures overlap and coordinate within an utterance [9]. Because it encodes temporal structure explicitly, the gestural score makes it possible to compare how the same symbolic unit is realized across languages.

Figure 2 summarizes cross-linguistic differences in gestural overlap that influence syllable perception, based on experimental schematics, illustrating how differences in gestural coordination yield different syllable count judgments. [4].

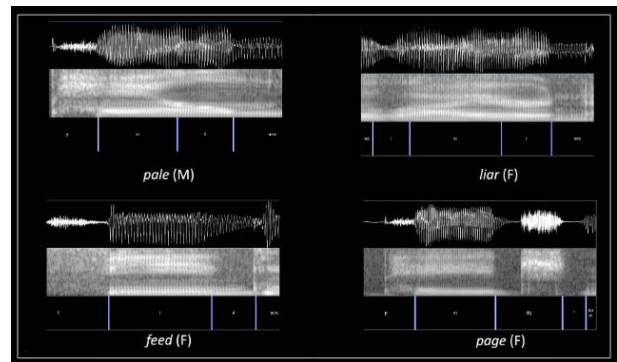


Figure 2. Gestural overlap patterns influencing syllable perception across languages.

Analyses of lexical tone in Serbian dialects, for example, show that tonal contrasts are associated with distinct patterns of articulatory timing rather than simple pitch targets [3]. Similarly, cross-linguistic experiments on syllable count judgments demonstrate that listeners are sensitive to gestural organization, not merely to the number of segments or vowels present [4].

5. Toward Gesture-Based Forced Alignment

5.1. Articulatory Features and Multilingual Modeling

Traditional forced alignment is often constrained by its reliance on acoustic features that vary significantly across languages. Early computational work on multilingual articulatory features

suggests that representations, such as place and manner of articulation, can provide a more language-general foundation for speech modeling than purely acoustic features [10]. By abstracting away from language-specific acoustics and focusing on shared articulatory dimensions, these models may offer a pathway toward more robust cross-linguistic alignment.

This shift toward Articulatory Phonology allows for the creation of a gestural score, which can visually demonstrate how even when two languages share the same IPA symbol, the actual physical realization, and thus the required alignment, is dictated by distinct temporal and spatial patterns of movement.

5.2. Recent Advances in Gesture-Informed Alignment

Recent research has begun to integrate gestural information directly into forced alignment workflows. [11] propose a joint framework for estimating articulatory movements, phoneme sequences, and alignments directly from speech, effectively linking the acoustic signal to underlying gestures. Similarly, SSDM adopts articulatory gestures as scalable units for modeling speech dysfluency, implicitly reframing alignment as a problem of gesture coordination rather than segment matching [12].

Figure 3 illustrates the architecture of gesture-informed alignment proposed by [11], highlighting the integration of articulatory estimation with phoneme alignment. Similarly, scalable speech dysfluency modeling adopts articulatory gestures as fundamental representational units [12].

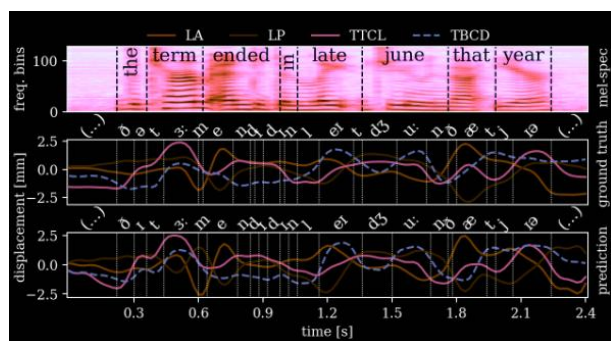


Figure C. *Gesture-informed alignment architecture integrating articulatory movement estimation with phoneme alignment.*

These approaches represent a conceptual shift: forced alignment is no longer merely the alignment of symbols to sound, but the alignment of physical actions to acoustic outcomes. This shift directly addresses the core limitation identified throughout the literature, that traditional FA systems struggle precisely because they ignore how speech is actually produced.

6. Conclusions

The review has argued that the persistent challenges faced by forced alignment in diverse languages arise from representational assumptions that overlook articulatory variation. While multilingual systems such as WebMAUS proved practical solutions for low-resource contexts, they remain constrained by segment-based models that conflate symbolic identity with phonetic equivalence. Articulatory phonology and the gestural score offer a theoretically grounded alternative, making explicit how identical IPA symbols can

correspond to distinct gestural realizations across languages. Emerging gesture-based alignment frameworks point toward a future in which forced alignment is not only computationally efficient but also linguistically informed and genuinely cross-linguistic.

7. References

- [1] T. Kempton, “Cross-language forced alignment to assist community-based linguistics for low resource languages,” in *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Honolulu: Association for Computational Linguistics, 2017, pp. 165–169. doi: 10.18653/v1/W17-0122.
- [2] C. Jones, “Evaluating cross-linguistic forced alignment of conversational data in north Australian Kriol, an under-resourced language,” *Lang. Doc.*, vol. 13, 2019.
- [3] R. Karlin, “1 Expanding the gestural model of lexical tone: evidence 2 from two dialects of Serbian”.
- [4] A. Popescu and I. Chitoran, “Linking gestural representations to syllable count judgments: A cross-language test,” *Lab. Phonol.*, vol. 13, no. 1, Oct. 2022, doi: 10.16995/labphon.7681.
- [5] T. Kislser, U. Reichel, and F. Schiel, “Multilingual processing of speech via web services,” *Comput. Speech Lang.*, vol. 45, pp. 326–347, Sept. 2017, doi: 10.1016/j.csl.2017.01.005.
- [6] J. Strunk, F. Schiel, and F. Seifart, “Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS”.
- [7] J. Al-Tamimi *et al.*, “A Romanization System and WebMAUS Aligner for Arabic Varieties”.
- [8] J. Leinonen, N. Partanen, S. Virpioja, and M. Kurimo, “Semiautomatic Speech Alignment for Under-Resourced Languages”.
- [9] “Browman_Goldstein_1989.”
- [10] S. Stuker, T. Schultz, F. Metze, and A. Waibell, “Multilingual articulatory features,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, Hong Kong, China: IEEE, 2003, p. I-144–I-147. doi: 10.1109/ICASSP.2003.1198737.
- [11] T. Weise *et al.*, “Speaker- and Text-Independent Estimation of Articulatory Movements and Phoneme Alignments from Speech,” July 03, 2024, *arXiv*: arXiv:2407.03132. doi: 10.48550/arXiv.2407.03132.
- [12] J. Lian *et al.*, “SSDM: Scalable Speech Dysfluency Modeling”.