Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment

Christian DiCanio,[a] Hosung Nam, and Douglas H. Whalen
Haskins Laboratories, 300 George Street, New Haven, Connecticut 06511

H. Timothy Bunnell
Center for Pediatric Auditory & Speech Science, Nemours Biomedical Research, 1600 Rockland Road, Wilmington, Delaware 19803

Jonathan D. Amith
Department of Anthropology, Gettysburg College, 300 North Washington Street, Gettysburg, Pennsylvania 17325-1400

Rey Castillo García
Secretaría de Educación Pública, Avenida de la Juventud, Chilpancingo, Guerrero, C.P. 39090, Mexico

(Received 7 October 2012; revised 26 June 2013; accepted 8 July 2013)

While efforts to document endangered languages have steadily increased, the phonetic analysis of endangered language data remains a challenge. The transcription of large documentation corpora is, by itself, a tremendous feat. Yet, the process of segmentation remains a bottleneck for research with data of this kind. This paper examines whether a speech processing tool, forced alignment, can facilitate the segmentation task for small data sets, even when the target language differs from the training language. The authors also examined whether a phone set with contextualization outperforms a more general one. The accuracy of two forced aligners trained on English (HMALIGN and P2FA) was assessed using corpus data from Yoloxóchitl Mixtec. Overall, agreement performance was relatively good, with accuracy at 70.9% within 30 ms for HMALIGN and 65.7% within 30 ms for P2FA. Segmental and tonal categories influenced accuracy as well. For instance, additional stop allophones in HMALIGN's phone set aided alignment accuracy. Agreement differences between aligners also corresponded closely with the types of data on which the aligners were trained. Overall, using existing alignment systems was found to have potential for making phonetic analysis of small corpora more efficient, with more allophonic phone sets providing better agreement than general ones.
 [http://dx.doi.org/10.1121/1.4816491]

PACS number(s): 43.72.Ar, 43.72.Lc, 43.70.Fq [CYE] Pages: 2235–2246

I. INTRODUCTION

Over the past thirty years, the documentation of endangered languages has become an urgent task for linguistic research. Part of what motivates this interest is the need to provide a more complete record of linguistic and cultural knowledge. Of the approximately 6900 languages now spoken, many are endangered (Krauss, 1992) and of these only a relatively small proportion are adequately documented. Indeed, it has been estimated that between 27 and 89% of all language families are threatened with complete disappearance (Whalen and Simons, 2012). Another motivating factor in language documentation is typological. Linguists still know relatively little about the diversity of language structures and patterns that exist. The development of any viable language typology or cross-linguistic theory of language production or perception must be able to account for the diversity of patterns that occur throughout the world, many of which are still undocumented.

Efforts at language documentation have been increased significantly in recent years due to the efforts of several funding agencies: the National Science Foundation Documenting Endangered Languages Program, the Endangered Language Fund, the Hans Rausing Endangered Language Program, Dokumentation Bedrohter Sprachen, and the Foundation for Endangered Languages. Hundreds of projects throughout the world collect language materials in endangered or minority languages and describe the grammatical structures of these languages. The collection of speech corpora for these languages comprises most of the available data for descriptive and exploratory linguistic analyses. While some of these data have contributed significantly to corpus-based grammatical research (see Bickel *et al.*, 2007 and Du Bois, 1987 for examples), their utility has been less evident with respect to phonological and phonetic description. Many documentation projects provide language materials in the form of corpora, lexicons, and grammar, with relatively little archived material specifically addressing phonetic or phonological questions. Given the nature of the data set for these endangered languages, can the material be made accessible to phonetic/phonological research?

One of the reasons for the absence of phonetic descriptions from corpora is the substantial time required for manual speech segmentation and annotation. Once proper transcription of speech corpora is completed by linguists and

[a]Author to whom correspondence should be addressed. Electronic mail: dicanio@haskins.yale.edu

speakers familiar with the language's structure, these transcriptions must be acoustically aligned with the recording itself. This segmentation process is even more time-consuming than transcription and it may also be prone to human errors (Jarifi *et al.*, 2008). Yet segmentation is a prerequisite for phonetic and phonological research on the documentation corpora.

Automatic methods of text-speech alignment (*forced alignment*) have been successful in the extraction of natural speech data for a variety of languages: well-studied languages such as English (Yuan and Liberman 2008, 2009), French (Adda-Decker and Snoeren, 2011), and Mandarin Chinese (Lin *et al.*, 2005) as well as less well-studied languages like Gaelic (Ní Chasaide *et al.*, 2006) and Xhosa (Roux and Visagie, 2007). Forced alignment, however, has not yet been applied to endangered language material produced by documentation projects. There are two main reasons for this. First, most of the research to develop forced alignment programs for a particular language is driven by commercial, not research interests. Corporations seek to build automatic systems that serve large languages with a large number of speakers. There is little commercial interest in devoting time to endangered languages, which by definition often have relatively few speakers. Second, forced alignment systems are often trained on large corpora (often between ten and hundreds of hours of speech) from a large number of speakers, e.g., between 40 (Malfrère *et al.*, 2003) and 630 speakers (Garofolo *et al.*, 1993). Even an automatic speech recognition system (ASR) built for under-resourced languages typically involves recordings from hundreds of speakers (Badenhorst *et al.*, 2011). Transcribed corpora from endangered language documentation projects typically come from a handful of speakers and typically consist of between 5–60 h of recordings. These data sets are limited in comparison with those that have been used for building forced alignment systems, so it is not clear that it would be feasible to build systems with current levels of data from most endangered language projects.

One way to take advantage of forced alignment for endangered language research would be to apply existing aligners trained on mainstream languages to the smaller corpora from different languages of interest. The viability of this approach has been tested for major languages, such as Hungarian, Russian, Czech, Malay, and English (Sim and Li 2008a, 2008b; Hai Do *et al.* 2012), but not for endangered language data from a documentation project. Another way to bring forced alignment to such languages is with multilingual speech recognition systems. Yet, while multilingual speech recognition has improved over the years, word error rates and misalignments still remain at very high levels (20–60%) (Boula de Mareüil *et al.*, 1999; Burget *et al.*, 2010; Dalsgaard *et al.*, 1991; Imseng *et al.*, 2012; Karafiát *et al.*, 2012). Better alignment has been found with language-dependent systems than with multilingual systems.

The current study examines how well two existing forced aligners work in the segmentation of endangered language data and specifically tests two hypotheses. The first is that forced alignment, even using a system trained on an unrelated language, can speed phonetic processing if the segment set is fairly similar between the target and training languages. The second is that a context-sensitive phone set will provide better phone mappings for under-resourced languages due to the greater likelihood that the broader set will allow for matches on finer acoustic detail than a narrower set. The data for the present test consist of an isolated word corpus from Yoloxóchitl Mixtec (YM, henceforth; ISO 639 code xty), an Oto-Manguean language spoken in a community in Guerrero, Mexico, about four hours southeast of Acapulco in the municipality of San Luis Acatlán. This language is practical for an English-based alignment system as the language's phoneme inventory is relatively simple and largely overlapping with that of English. The corpus consists of six repetitions of 261 YM words spoken by ten native speakers (15 660 tokens), totaling approximately 7 h of speech. The results of two different forced aligners were compared with hand-labeling. The first was P2FA, developed at the University of Pennsylvania Phonetics Laboratory (Yuan and Liberman, 2008). The second was the Nemours SRL aligner (henceforth, HMALIGN), developed in the Speech Research Laboratory at the Nemours Alfred I. duPont Hospital for Children by Timothy Bunnell and colleagues. HMALIGN is a stand-alone version of the aligner developed for the ModelTalker TTS system voice recording program (Yarrington *et al.*, 2008). The viability of forced alignment for this type of corpus and the relative accuracy of the two phone sets were evaluated by comparing the agreement of boundary locations between forced alignment and hand-labeling. Agreement was compared both generally and for different phonological classes. Statistical analyses were performed to test how well each aligner did for each phonological class. The results here are novel in both their application to endangered language data and in their relevance to speech technology. To date, we know of no such study that examines the temporal aspects of alignment using an aligner trained on a different source language from that which is tested. The current study seeks to fill this gap.

A. Forced alignment

Progress in speech processing has made it possible to automatically segment transcribed speech. Constructing speech processing systems for a language usually implies developing an ASR. This requires a significant amount of time-coded and transcribed speech data along with the corresponding text and pronunciation dictionary (Livescu *et al.*, 2012; Rabiner *et al.*, 1993). Unless the training corpus is read speech based on a script, researchers have usually had to provide transcriptions by hand. Though, recent advances in deep neural networks have shown automatically generated transcripts to be very useful in training speech systems (Hinton *et al.*, 2012). Once there is enough annotated data to build a successful ASR system, that system can be used to automatically mark phonetic segments in a corpus of transcribed utterances. The development of speech corpora has, in fact, relied on forced alignment to provide a first pass segmentation which is subsequently corrected by hand (Zue and Seneff, 1990; Zue *et al.*, 1996). The system does this by deriving word- and phone-level labeling from a word-level

transcription and a "dictionary" (a lexical list from the corpus). For example, if a forced aligner is provided with the sound file and transcription of "bad," it will look up the phones for the word and generate the temporal boundaries for the word and the phones (/b/, /æ/, /d/). Misaligned boundaries are then repaired and used to refit the forced alignment model parameters.

When developing a speech recognizer for a new language, it is also possible to use a forced alignment built for a *different* language with a well-trained recognizer to provide phone segmentation. This is especially useful for under-resourced languages lacking a substantial set of training data for a complete ASR system (Badenhorst *et al.*, 2011; Livescu *et al.*, 2012; Sim and Li, 2008a, 2008b). The accuracy of this strategy depends closely on the fit between the phone set used to build the aligner and the phone set of the target language. In a series of studies Sim and Li (2008a, 2008b) found that a Hungarian phone recognizer performed better than a Russian one in recognizing Czech phones. This finding is somewhat surprising considering that Russian and Czech are Slavic languages (Indo-European), while Hungarian is Ugric (Uralic), belonging to a completely separate language stock. Typically, aligner performance improves by including training data from genetically similar languages, such as Dutch, English, and German training data for Afrikaans (Imseng *et al.*, 2012). However, the consonant systems of Czech and Hungarian are quite similar (especially among obstruents), while Russian has a distinct set of palatalized consonants which are missing in Czech (Dankovičová, 1997; Jones and Ward, 1969). Thus, a closer correspondence between the phone inventory of the training and target languages is a more accurate predictor of the accuracy of forced alignment than the genetic similarity of the languages.

Another way of achieving a closer correspondence between phone inventories is by including context-sensitive phones. Sim and Li (2008a, 2008b) found that including contextual information in phone mapping resulted in a relative improvement of 5 to 7%. By their nature, certain monophone sets contain more contextual information than others. In this study, we compare the accuracy of two English-trained forced aligners (P2FA and HMALIGN) on a elicited corpus of YM speech. P2FA's acoustic models are GMM-based monophone-HMMs trained using the SCOTUS corpus (Garofolo *et al.*, 1993), which contains oral arguments made before the Supreme Court of the United States. Each HMM state consists of 32 Gaussian mixture components on 39 perceptual linear predictive coefficients (12 cepstral plus energy, delta, and acceleration). P2FA employs CMU phones, which do not show allophonic variants in English. These phones are context-independent. On the other hand, HMALIGN uses a set of discrete monophone HMMs trained on data from the TIMIT corpus (Garofolo *et al.*, 1993). For HMALIGN, separate 375-word codebooks were trained on vectors of eight mel frequency cepstral coefficients, plus their delta and acceleration coefficients. It employs the larger Applied Science and Engineering Laboratories (ASEL) extended English phone set, which includes some allophones, such as aspirated and unaspirated stop variants. While these aligners differ from each other in a number of ways, the inclusion of some allophones in HMALIGN allows us to examine their role in aligner agreement for endangered language data. The phone-phoneme mappings are shown in Table IV and described in Sec. II B.

TABLE I. Consonant inventory in Yoloxóchitl Mixtec.

	Bilabial	Dental	Alveo-palatal	Palatal	Velar	Labialized velar	Glottal
Stop	p[a]	t			k	k^w	ʔ
Pre-nasalized stop	^{m}b	^{n}d			^{ŋ}g		
Fricative	β	s	ʃ				
Affricate			tʃ				
Nasal	m	n					
Approximant				j		w	
Lateral approximant		l					
Tap		ɾ					

[a]/p/ is quite rare in native vocabulary.

B. Language background

Yoloxóchitl Mixtec is an Oto-Manguean language spoken in the state of Guerrero, Mexico. The segmental inventory is rather simple, but the lexical tone inventory is quite complex. The consonant and vowel inventories are given in Tables I and II, respectively.

The syllable structure in YM is simple, consisting only of CV syllables. While the glottal stop appears in the inventory above, it is best considered a prosodic characteristic of words which occurs in foot-medially. Like many Mixtec languages, all content words are minimally bimoraic (Macken and Salmons, 1997), consisting either of a consonant followed by a long vowel (CVV) or two syllables with a short vowel (CVCV), and maximally trimoraic, e.g., CVCVV or CVCVCV. While the tonal morphophonology is quite complex, words are morphologically simple. Verbs may be preceded by a single prefix morpheme (usually marking aspect). All words may be followed by one of several personal pronominal enclitics, most of the shape /CV/.

The tonal patterns in YM, occurring as they do primarily on the vocalic segments, do not need a separate alignment. Although they are not directly represented, however, the tones may influence the alignment because they impose un-English-like patterns on F_0 and amplitude. Thus, a brief inventory is provided here. Lexically, there are four level tones and five contour tones consisting of two levels. Up to five different tones may occur on the initial mora of a root and up to eight may occur on the second mora. The distribution of tone varies in relation to both word size and glottalization (for example, rising or falling tones almost never occur in the second mora of a word with a glottalized vowel in the

TABLE II. Vowel inventory in Yoloxóchitl Mixtec.

	Front	Central	Back
Close	i, ĩ		u, ũ
Close-mid	e, ẽ		o, õ
Open		a, ã	

TABLE III. Tones in Yoloxóchitl Mixtec (1 is low, 4 is high).

	Tone—Initial mora				
Tone—Final mora	/1/	/3/	/4/	/13/	/14/
/1/	ja^1a^1 "slow"		$ja\text{ʔ}^4a^1$ "grey"		$na\text{ʔ}^{14}a^1$ "demonic"
/2/		$j\tilde{u}^3\tilde{u}^2$ "town"	$\int a\text{ʔ}^4a^2$ "cooked maize"	$t\tilde{i}^{13}\text{ʔ}\tilde{i}^2$ "yielded"	$\int\tilde{a}\text{ʔ}^{14}\tilde{a}^2$ "greasy"
/3/	ta^1a^3 "man"	$^nde^3e^3$ "face down"	βi^4ka^3 "rich"		$nu^{14}u^3$ "face"
/4/	$\int i^1i^4$ "grandfather"	na^3a^4 "night"	ja^4a^4 "cold (personality)"		$je\text{ʔ}^{14}e^4$ "door"
/13/			$t\int e^4e^{13}$ "large"		
/24/			ja^4a^{24} "tongue"		
/32/	$\int a^1ko^{32}$ "opossum"				
/42/	$ta^1k^wi^{42}$ "water"	$j\tilde{u}^3\tilde{u}^{42}$ "dark"			

first mora). These aspects of the tonal distribution are not discussed here. Table III shows the YM tonal inventory with representative bimoraic monomorphemic words. Up to 20 tonal combinations have been documented. In morphologically complex words, more combinations are possible as certain morphemes may be marked with tone.

Two additional phonological patterns in YM are relevant to the current study. First, there is a robust process of progressive vowel nasalization. The vowels which follows the nasal consonant in the same syllable is phonetically nasalized, e.g., /na^3a^4/ "night" is phonetically [$\tilde{na}^3\tilde{a}^4$]. This process does not affect vowels following prenasalized stops (e.g., [nd]). As a result of this process, there is no contrast in vowel nasalization after nasal consonants, i.e., all vowels following a nasal consonant are nasalized. English vowels can be allophonically nasalized, but nasalization occurs on the vowel preceding a nasal consonant (Beddor and Krakow, 1999). Thus both the phonemic nasalization and the direction of allophonic nasalization in YM differ from that in English. Second, the glottal stop occurs only word-medially, either intervocalically, e.g., /$ja\text{ʔ}^4a^1$/ "gray," or preceding a sonorant consonant, e.g., /$sa\text{ʔ}^3ma^4$/ "cloth to wrap tortillas." It is a feature of the phonological foot in YM and it frequently surfaces as creaky phonation in both contexts where it occurs. This pattern does not have a direct analog in English.

II. METHODS

A. Data collection and hand-labeling

The data set for the current study comes from a corpus of isolated YM words collected in 2010 by J. Amith, C. DiCanio, and R. Castillo García. It comprises 261 words, repeated six times, produced by each of ten native speakers of YM. A total of 15 660 words were recorded. The primary purpose of this elicitation was to explore differences in the production of tone in words of different sizes. Of the 261 words, 169 were disyllabic, 89 were monosyllabic, and three were trisyllabic. Including morphologically derived words, a total of 30 different tonal melodies occurred on these words. For most of these, the onset consonants were balanced for voicing. Five of the 30 tonal melodies were quite rare and only a single word was recorded for each of these. The entire corpus was hand-labeled by L. DiDomenico, a linguistics graduate student at Université Lyon 2 and manually checked by the first author.

B. Data coding and forced alignment

The YM data is phonologically different from that of English, the language on which each of the forced aligners is trained. Despite such differences, care was taken to select the closest possible phone correspondence in the forced alignment. The phone correspondences are shown in Table IV. For both P2FA and HMALIGN, certain phonological categories were neutralized during alignment. In particular, nasal and oral vowels were treated as oral phones, as the aligners were

TABLE IV. Phone correspondences with YM phonemes.

Mixtec	P2FA	HMALIGN
/p/ [p]	P [p^h, p]	PP [p]
/t/ [t]	T [t^h, t, $\widehat{t\text{ʔ}}$, ɾ]	TT [t]
/k/ [k]	K [k^h, k]	KK [k]
/k^w/ [k^w]	K [k^h, k]	KK [k]
/ʔ/ [ʔ]	T [t^h, t, $\widehat{t\text{ʔ}}$, ɾ]	TQ [$\widehat{t\text{ʔ}}$]
/^{n}d/ [^{n}d]	N [n]	NN [n]
/tʃ/ [tʃ]	CH [tʃ]	CH [tʃ]
/m/ [m]	M [m]	MM [m]
/n/ [n]	N [n]	NN [n]
/β/ [β, β̞, b]	W [w]	WW [w]
/s/ [s]	S [s]	SS [s]
/ʃ/ [ʃ]	SH [ʃ]	SH [ʃ]
/ɾ/ [ɾ]	R [ɻ, ɹ]	RR [ɻ, ɹ]
/l/ [l]	L [l, ɬ]	LL [l, ɬ]
/j/ [j]	Y [j]	JY [j]
/i/ [i]	IY [i, ĩ]	II [i, ĩ]
/ĩ/ [ĩ]	IY [i, ĩ]	II [i, ĩ]
/e/ [e, ɛ]	EH [ɛ, ɛ̃]	EH [ɛ, ɛ̃]
/ẽ/ [ẽ, ɛ̃]	EH [ɛ, ɛ̃]	EH [ɛ, ɛ̃]
/a/ [a]	AA [ɑ, ã, a, ã]	AA [ɑ, ã, a, ã]
/ã/ [ã]	AA [ɑ, ã, a, ã]	AA [ɑ, ã, a, ã]
/o/ [o, ɔ]	AO [ɔ, ɔ̃]	AO [ɔ, ɔ̃]
/õ/ [õ, ɔ̃]	AO [ɔ, ɔ̃]	AO [ɔ, ɔ̃]
/u/ [u]	UW [u, ũ, ʉ, ʉ̃]	UW [u, ũ, ʉ, ʉ̃]
/ũ/ [ũ]	UW [u, ũ, ʉ, ʉ̃]	UW [u, ũ, ʉ, ʉ̃]

trained on English data; the substantial coarticulatory vowel nasalization (Beddor and Krakow, 1999) was ignored by the systems. Moreover, two pairs of phonemes were neutralized in alignment. Both /k/ and /k^w/ were treated as the same phone, /k/, and both /n/ and /^{n}d/ were treated as the same phone, /n/. Note that collapsing these categories does not entail the loss of detail in the transcription itself. Those categories were simply aligned with a phonetic segment that the English-based systems have as a phone. The duration of both of these complex segments was more similar to a simple segment in English than to a sequence (i.e., /kw/ and /nd/). It was assumed that the correspondence chosen would be optimal, but the alternative was not tested.

During a pilot phase, different phone variants were tried for two other phonological contrasts in YM: /β/ and /ʔ/. The voiced bilabial fricative frequently varies in its production between a voiced frictionless bilabial continuant β, a voiced bilabial fricative [β], and a voiced bilabial stop [b]. A /b/ phone was used during the pilot phase, but agreement was quite poor. A /w/ phone provided a much better match for this YM phoneme. The glottal stop is frequently lenited or produced as creaky phonation overlaid on the adjacent vowels and consonants. An /h/ phone was used during the pilot phase for this, but agreement was found to improve with a /t/ phone. While English does not have phonologically contrastive glottalization, a frequent realization of coda /t/ contains creaky phonation, e.g. [t] (Huffman, 2005). Thus, there is a high degree of similarity between YM and English sequences with pre-consonantal glottalization. For instance, compare English "chutney" ['tʃʊt̰ʔ.ni] with YM "kill(s)" [ʃaʔ^4.ni^{24}].

The phone set for HMALIGN differs from that of P2FA as the former contains separate stop allophones for voiceless unaspirated English phones [p, t, k] (PP, TT, KK), voiceless aspirated English phones [p^h, t^h, k^h] (HP, HT, HK), and the glottalized English phone [t̰ʔ] (TQ). As YM contains only unaspirated voiceless stops, these phones from HMALIGN were used. YM glottal stops were coded with the TQ phone from HMALIGN (rather than the [t] of P2FA).

We ran the forced aligners as follows: First, each forced aligner took a sequence of words and the associated speech utterance as an input. In each model, a pronunciation dictionary was constructed where the pronunciations of the YM word were coded using the phone set specific to that model. Table IV shows the phonemic inventory of YM and phone correspondences for each aligner, along with its phonetic exponents in brackets. For example, a YM word, /^{n}du^{4}βa^2/ "he/she falls backward" is coded as /NN UW1 WW AA1/ in HMALIGN and /N UW1 W AA1/ in P2FA. Note that all vowels were coded as stressed (marked with "1" after each vowel). During the pilot phase, unstressed vowel phones were used. This resulted in substantial loss of agreement. Agreement improved when all vowels were treated as equally "stressed." Unlike English, Mixtec does not have substantial vowel reduction despite having differences in stress. Second, the YM speech data were downsampled from 48 to 16 kHz for both P2FA and HMALIGN to achieve parity between the aligners' training data and the Mixtec test data. To compare hand-labeling and results from the two models with one another, we needed to align all the datasets (the hand labeling, the P2FA labeling, and the HMALIGN labeling) together. We excluded tokens with missing phones and extraneous pauses. In total, 5 of the 261 words were excluded from all speakers. The extraneous pauses came entirely from short pauses (sp) that were inserted by P2FA. In total, 5232 short pauses were inserted out of 83 768 alignments, representing 6.2% of the data. HMALIGN did not insert any short pauses. Both aligners placed boundaries with 10 ms temporal resolution.

C. Statistical measures of aligner performance

Start and end points for each phonemic interval were extracted from the hand-labeled and the force aligned textgrid files using a script written for PRAAT (Boersma and Weenink, 2012). The relative differences between the segment boundaries for the hand-labeled files and the force-aligned files were compared. This comprised a total of 54 540 comparable segments. In addition to the temporal data, all words were coded for the presence of glottalization, vowel nasalization, size (monosyllabic/disyllabic), and tonal category. These categories were used to organize the data and to test how agreement was influenced by the phonological contrasts in the language. Boundary agreement and statistical tests were analyzed using R (R Development Core Team, 2012). Statistical tests were run for each phonological category, corresponding to the separate result sections below. In each test, lexical items were treated as a random effect in a linear mixed effects model and the choice of aligner (P2FA or HMALIGN) was treated as a fixed effect. In mixed effects models, *p*-values are calculated not based on a standard degrees of freedom value, but on the upper bound of the degrees of freedom (=total observations − fixed effects parameters). This is typical for mixed effects models, as the estimation of the degrees of freedom is not clearly established (Baayen, 2008; Bates, 2005). Two sets of *p*-values were obtained from the mixed effects model, one using Markov-chain Monte-Carlo (MCMC) sampling and another based on the *t*-distribution. The *p*-values reported here derive from the *t* distribution, but were validated against those from the MCMC simulation, which adjusts for random effects. The value given with the *t*-statistic, e.g., *t*[num], reflects the upper bound on the degrees of freedom.

III. AGREEMENT AND ALIGNER RESULTS

A. Results 1: General alignment accuracy

Agreement was fairly good for both aligners. Agreement of HMALIGN was 70.9% in 30 ms compared to 65.7% in 30 ms for P2FA. Table V shows agreement at different thresholds.

TABLE V. Agreement with hand-labeling.

Threshold	P2FA	HMALIGN
10 ms	32.3%	40.6%
20 ms	52.3%	61.4%
30 ms	65.7%	70.9%
40 ms	74.8%	81.2%
50 ms	79.6%	86.7%

Agreement for HMALIGN was better than for P2FA. The results of a linear mixed effects model showed a strong effect of aligner on agreement, both at start points ($t[142734] = 6.2$, $p < 0.001$) and at endpoints ($t[142734] = 6.0$, $p < 0.001$). As stated, agreement of HMALIGN was 70.9% in 30 ms compared to 65.7% in 30 ms for P2FA. This reflects a 15.2% error reduction between the models. Error reduction reflects the quotient of the error differences divided by the larger error value, e.g., ((100–65.7)–(100–70.9))/(100–65.7).

Agreement for both aligners was low in comparison with forced alignment on models trained on their target language, which typically average above 80% within 20 ms (Hosom, 2009; Malfrère *et al.*, 2003). This pattern is expected though, as the alignment systems were not trained on YM data. Nevertheless, the agreement levels reported in Table V hide some important differences in how different sound types and the position of a segment in a word influence forced alignment. Figure 1 shows agreement and segmental start and end points for consonants and vowels for each forced aligner. Note that while the median agreement in Fig. 1 is centered around zero (indicating overall good agreement), we are particularly interested in the amount of error in alignment, represented by the larger quartiles in the boxplot.

Figure 1 shows greater variability in agreement for consonants at starting points than at endpoints and for vowels at endpoints than at starting points. Agreement at a consonant starting point corresponds to both #-C and V-C transitions, while agreement at a consonant endpoint corresponds only to C-V transitions (as YM has no codas). Agreement at a vowel starting point corresponds only to the C-V transition, while agreement at a vowel endpoint corresponds to both the V-C and the V-# transition. Generally speaking, the acoustic cues to the C-V transition are stronger than those at the V-C transition (Ohala, 1990; Manuel, 1991; Hura *et al.* 1992). This pattern influences the alignment measurements. Further, the boundary between a vowel and subsequent silence is somewhat arbitrary and differences between automatic and hand-aligned data for those boundaries may reflect sensitivity to different cues with human labelers; humans may be more sensitive to amplitude and voicing features, while automatic aligners may weigh spectral shape features more heavily.

Given that the corpus consisted solely of words spoken in isolation, many segments had to be aligned at a word-boundary. The acoustic cues separating silence from these boundaries may not be sufficiently strong for precise estimates of alignment. Such conditions occur when, for instance, a voiceless stop consonant is the word-onset or if a speaker gradually devoices a word-final vowel. A linear mixed-effects model was used to examine the effect of these boundaries on agreement, with boundary and aligner treated as independent variables. Boundaries were coded as *edge* or *non-edge*. Agreement at edges corresponded to either endpoint agreement for final vowels or start point agreement for initial consonants. Agreement at non-edges corresponded to start point agreement in the remainder of the cases. Both main effects were significant. The data in Table VI show differences in agreement at word edges and non-edges. Agreement at non-edges was significantly better than agreement at edges ($t[85728] = 3.8$, $p < 0.001$). At 30 ms, the average agreement at edges was 63.8% for P2FA and 54.0% for HMALIGN. The average agreement at non-edges was 72.3% for P2FA and 81.5% for HMALIGN. These values reflect an error reduction of 23.4% for P2FA and 59.8% for HMALIGN. Agreement for HMALIGN was significantly better than for P2FA ($t[85728] = 10.6$, $p < 0.001$).

There was a significant interaction between the forced aligner and boundary type. Agreement was higher for P2FA than HMALIGN at word boundaries (21.3% error reduction) but higher for HMALIGN than P2FA at non-boundaries (33.3% error reduction) ($t[85728] = 10.2$, $p < 0.001$).

TABLE VI. Agreement with hand-labeling across positions.

	P2FA by position				HMALIGN by position			
Threshold	#-C	C-V	V-C	V-#	#-C	C-V	V-C	V-#
10 ms	11.9%	59.4%	40.6%	48.1%	23.2%	54.6%	46.0%	33.6%
20 ms	35.7%	70.3%	59.6%	54.7%	43.5%	78.6%	68.9%	40.8%
30 ms	64.3%	77.6%	67.9%	63.2%	58.1%	85.3%	77.6%	49.9%
40 ms	84.6%	81.5%	73.0%	71.7%	82.6%	89.0%	83.2%	59.6%
50 ms	91.8%	84.3%	76.9%	77.4%	92.2%	91.3%	87.0%	69.2%

B. Results 2: Category-specific effects on alignment accuracy

Just as the position of a segment within the word influences agreement for each of the forced aligners, phonological

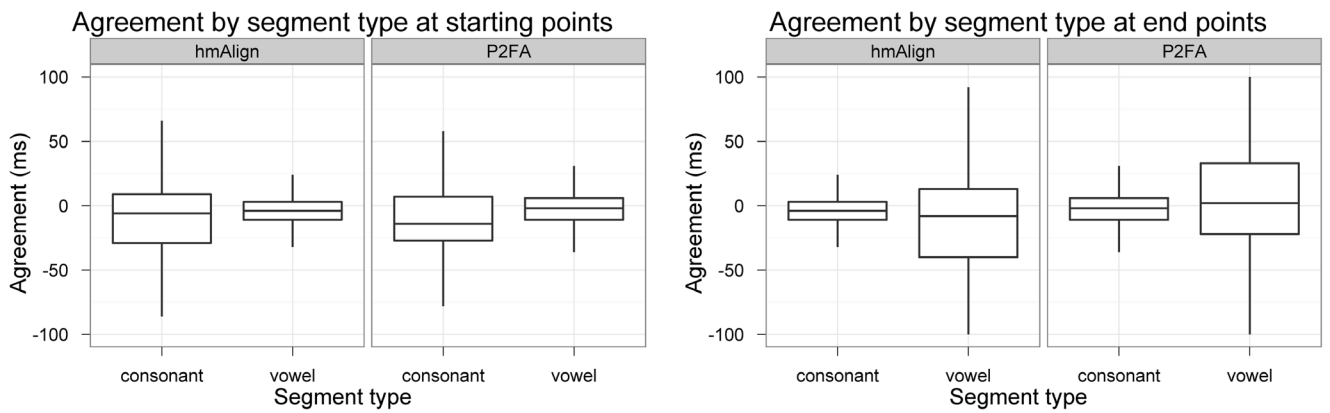

FIG. 1. Agreement for consonants and vowels across aligners.

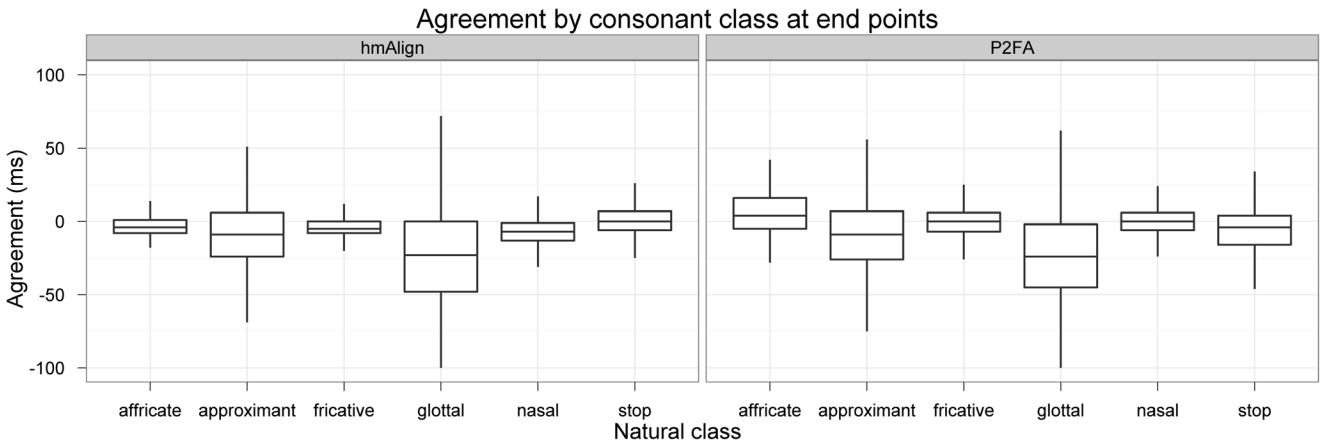

FIG. 2. Agreement for consonant classes across aligners.

classes also differ in agreement. Three types of classes were considered here: Consonantal phonological classes (e.g., stop, fricative, nasal); vowel quality and nasalization; and tone.

1. Consonantal phonological class

Consonant data was organized into six different natural classes: stops, fricatives, affricates, glottal stops, nasals, and approximants, and analyzed using a linear mixed effects model with natural class and aligner as independent variables. Neither of the main effects were significant, but agreement for stops was near significance ($t[56916] = 1.9$, $p = 0.06$). Figure 2 shows agreement for both aligners for each consonant class and Table VII shows agreement at different thresholds.

A significant aligner x stop interaction occurred as well ($t[56916] = 3.6$, $p < 0.001$). This reflected lower error in agreement for stops with HMALIGN than with P2FA. In general, stops fricatives, affricates, and approximants showed better agreement with HMALIGN, with an average agreement of 89.0% vs 81.6% at 30 ms. This reflects an error reduction of 40.2% between aligners. Agreement was slightly better for nasals with P2FA (92.0% at 30 ms) than with HMALIGN (89.3% at 30 ms). This reflects an error reduction of 25.2%. Overall, agreement was worst for glottal stops; these are considered in more detail in Sec. III C 2.

TABLE VII. Agreement with hand-labeling across natural classes at endpoints.

	Threshold	Stop	Fricative	Affricate	Nasal	Approximant
P2FA	10 ms	42.3%	65.6%	43.5%	65.2%	26.1%
	20 ms	59.0%	91.5%	77.6%	87.4%	50.4%
	30 ms	65.2%	98.3%	94.6%	92.0%	68.1%
	40 ms	69.8%	99.2%	98.8%	93.2%	80.2%
	50 ms	73.7%	99.3%	99.1%	94.2%	87.1%
HMALIGN	10 ms	63.1%	77.5%	77.1%	48.9%	28.3%
	20 ms	78.4%	99.0%	98.8%	83.3%	53.5%
	30 ms	83.6%	99.6%	99.3%	89.3%	73.3%
	40 ms	87.3%	99.7%	99.8%	90.6%	86.1%
	50 ms	90.2%	99.8%	100%	91.5%	91.4%

2. Vowel quality and nasalization

Vowels were marked for quality (i, e, a, u) and for nasality (nasal or oral). While nasal vowels were not distinguished from oral vowels during alignment, this categorization allowed a test of whether vowel nasality affected alignment accuracy; if collapsing across the two kinds of nasality was indeed benign, there should be no difference in those measurements. For vowel quality, agreement at start points and endpoints was analyzed using a linear mixed effects model with vowel quality and aligner as independent variables. There was a significant effect of the aligner on agreement for all vowels ($t[52146] = 7.4$, $p < 0.001$). For all vowels except /i/, agreement was higher with HMALIGN than with P2FA. The average agreement for vowels is 73.1% at 30 ms with P2FA and 80.5% at 30 ms with HMALIGN. This reflects an error reduction of 27.5% between aligners. A significant general effect of vowel quality on agreement was also found for /i/ ($t[52146] = 2.3$, $p < 0.05$). For both aligners, agreement was better for /i/ than for the other vowels, with the exception of /a/ for HMALIGN. Figure 3 shows agreement for each vowel quality for each aligner. Agreement values at different thresholds for each aligner are given in Table VIII.

In order to examine agreement for nasal and oral vowels, a linear mixed effects model with vowel nasality and aligner as independent variables was run. No significant effect of

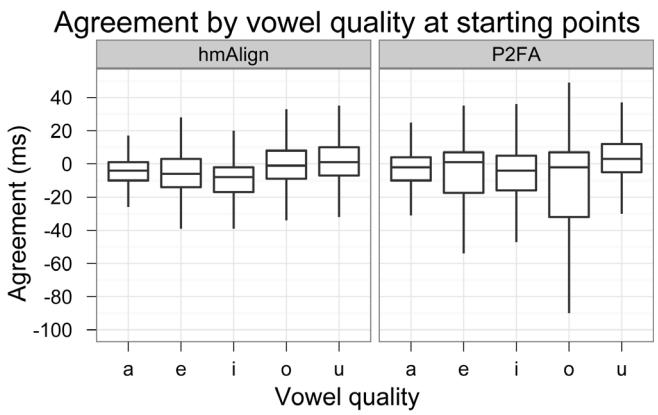

FIG. 3. Agreement for vowel qualities across aligners.

TABLE VIII. Agreement with hand-labeling across vowel qualities at start points.

	Threshold	/i/	/e/	/a/	/o/	/u/
P2FA	10 ms	47.2%	40.2%	54.5%	41.1%	42.5%
	20 ms	71.5%	59.1%	74.2%	60.3%	62.2%
	30 ms	82.7%	66.8%	80.4%	64.5%	70.9%
	40 ms	88.2%	70.4%	84.0%	67.4%	74.5%
	50 ms	91.8%	73.5%	86.4%	70.9%	77.0%
HMALIGN	10 ms	42.6%	43.0%	61.9%	49.0%	47.7%
	20 ms	70.8%	67.0%	85.4%	71.6%	70.0%
	30 ms	80.7%	73.5%	90.4%	78.8%	78.9%
	40 ms	86.3%	79.0%	93.0%	82.9%	83.3%
	50 ms	89.5%	83.0%	94.6%	86.8%	86.4%

vowel nasality on agreement was found. A significant effect of aligner on agreement with vowel nasality was found ($t[28030] = 3.9$, $p < 0.001$). For HMALIGN, agreement for nasal vowels was worse (79.1% at 30 ms) than for oral vowels (87.0% at 30 ms). The opposite pattern occurred for P2FA; agreement was more accurate for nasal vowels. Figure 4 shows agreement for each aligner with nasal vowels. Table IX shows agreement values at different thresholds.

For oral vowels, agreement was higher with HMALIGN than with P2FA. Agreement was 87.0% at 30 ms with HMALIGN but 74.8% at 30 ms with P2FA. This reflects an error reduction of 48.4%. However, this finding is correlated with the higher agreement values for nasal consonants with P2FA discussed in the previous section. Due to the process of progressive nasalization (see Sec. I B), vowels following a nasal consonant are nasalized in YM. Thus, agreement of vowels at starting points following nasal consonants was, as expected, higher for P2FA than for HMALIGN.

3. Tone

Generally speaking, a language's word-level prosody is not considered as a factor in forced alignment. Tone may be explicitly used as a first step in alignment of tone languages (Lei *et al.*, 2005), but there is no evidence that excluding it would cause an increase in error. Yet, there are reasons to predict that tone or lexical stress may influence segment-level forced alignment. There is a relationship between the amplitude envelope of the acoustic signal and F_0 (Whalen and Xu, 1992). Thus, it stands to reason that abrupt changes in amplitude, correlated with F_0 changes, can result in misalignment, especially if the forced aligner relies heavily on such amplitude perturbations.

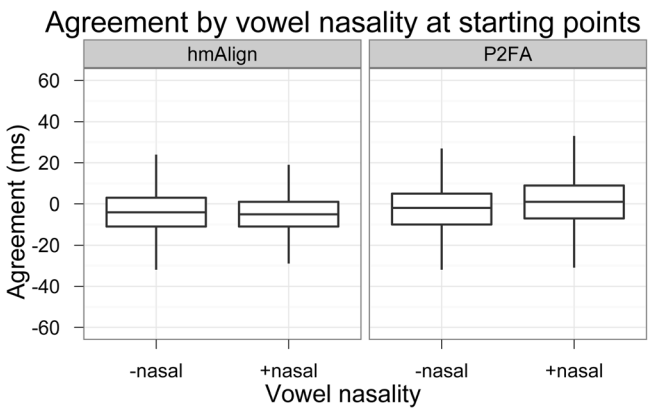

FIG. 4. Agreement for vowel nasalization across aligners.

TABLE IX. Agreement with hand-labeling for nasal vowels in disyllabic words.

	P2FA		HMALIGN	
Threshold	Oral	Nasal	Oral	Nasal
10 ms	45.1%	55.2%	54.3%	47.6%
20 ms	65.3%	80.9%	77.9%	74.1%
30 ms	74.8%	86.9%	87.0%	79.1%
40 ms	80.8%	88.2%	92.3%	81.3%
50 ms	84.8%	89.3%	95.0%	82.8%

The tonal inventory of YM is quite large in comparison to other tonal languages (Yip, 2002), if we treat the syllable as the tonal domain. If we treat the mora as the tonal domain, the inventory is much simpler. Current practice is to treat the syllable as the domain of tone. Rather than having the large number of categories where each tone is treated separately, each tone was placed into a tonal shape category: level, falling, rising, concave, convex, double rise (e.g., 14.14, with the dot dividing the moras), and double fall (e.g., 31.32). Agreement was tested in a linear mixed effects model with tonal category and aligner as the independent variables. The results showed no main effects in agreement for specific tonal categories, however, there was a significant interaction between tonal category and aligner for P2FA. Agreement was lower for rising tones and double rising tones with P2FA than with HMALIGN ($t[51834] = 2.3$, $p < 0.05$, $t[51834] = 2.3$, $p < 0.05$, respectively).

C. Results 3: Phoneme-specific effects on alignment accuracy

Certain language-specific phones may cause particular problems for forced alignment. In YM, two phonological contrasts are notably different from the English data on which each aligner was trained: unaspirated stops and glottalization. Each of these contrasts was considered separately in order to highlight how forced alignment performs with particular language-specific patterns.

1. Unaspirated stops

YM has four unaspirated stops (/p, t, k, k^w/) though only two categories were used for the purposes of alignment (/t, k/). The /p/ phoneme is quite rare in native words and only occurred on one word in the corpus, so it was excluded. The /k^w/ phoneme was more common, but it could not be easily matched to a single phone in the phone set for either aligner, so it was coded as /k/. Agreement of the two stop phones was examined with a linear mixed effects model with the specific phoneme label and the aligner as the independent variables. Results for both

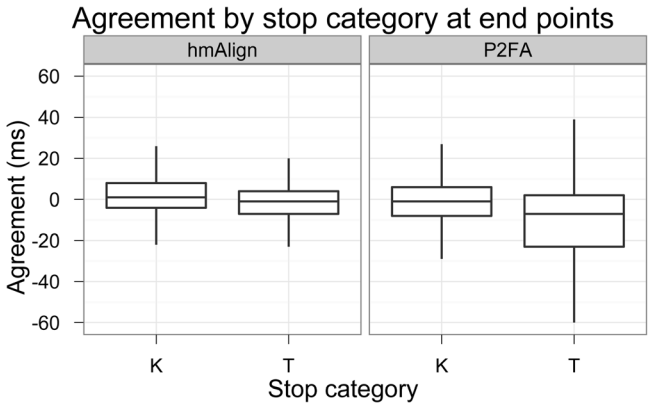

FIG. 5. Agreement for stops across aligners.

forced aligners are shown in Fig. 5. Agreement values are given in Table X.

There was a significant main effect of Aligner on stop agreement ($t[19410] = 5.3$, $p < 0.001$). Agreement for stops was higher for HMALIGN (91.5% at 30 ms) than for P2FA (68.5% at 30 ms). This reflects an error reduction of 25.1% across aligners. There was also a significant interaction between aligner and agreement for /t/ compared to /k/ for P2FA ($t[19410] = 2.3$, $p < 0.05$). Agreement for /t/ was substantially worse (61.3% at 30 ms) than for /k/ (75.6% at 30 ms). Agreement for /t/ was substantially better with HMALIGN (98.0% at 30 ms) than for P2FA, with an error reduction of 37.4%. An explanation for this pattern is provided in Sec. IV.

2. Positional effects on glottal stop agreement

Glottalization in YM occurs in two contexts, intervocalically (V?V) and before a sonorant consonant in word-medial position (V?NV). The latter position is akin to allophonic American English /t/, but the former is unlike any typical pattern in American English phonology, which should result in lower agreement. For each aligner, agreement at start points and endpoints for glottal stops was assessed using a linear mixed effects model with position and aligner as independent variables. Figure 6 shows agreement at each position for each aligner.

Both main effects were significant. There was a significant effect of the aligner on agreement for glottal stops ($t[6642] = 4.2$, $p < 0.001$). Agreement was better with HMALIGN (33.2% at 30 ms) than with P2FA (10.9% at 30 ms). There was also a significant main effect of position on agreement ($t[6642] = 4.9$, $p < 0.001$). Agreement was worse for intervocalic glottalization than for pre-consonantal glottalization. Finally, there was a significant interaction between the aligner and position ($t[6642] = 9.6$, $p < 0.001$). Agreement for glottal stops in the /V?V/ context was substantially worse for P2FA (10.3% at 30 ms), than for HMALIGN (35.7% at 30 ms). This reflects an error reduction of 71.1% in HMALIGN. Agreement for glottal stops in the /V?CV/ context was 11.5% at 30 ms for P2FA, but 30.8% at 30 ms for HMALIGN. This reflects an error reduction of 62.6% with HMALIGN. Table XI shows agreement values at different thresholds for each aligner.

TABLE X. Agreement with hand-labeling for stops across aligners.

	P2FA		HMALIGN	
Threshold	/t/	/k/	/t/	/k/
10 ms	40.7%	52.5%	83.5%	62.6%
20 ms	53.8%	71.3%	97.0%	79.2%
30 ms	61.3%	75.6%	98.0%	84.9%
40 ms	68.1%	77.9%	98.3%	88.3%
50 ms	73.3%	79.6%	98.5%	91.3%

IV. DISCUSSION

A. The utility of forced alignment for language documentation corpora

The present results indicate that a sizable majority of boundaries are within 30 ms of hand labeling for our data set. This means that automatic alignments should be a useful beginning point for labeling a new dataset. In an ideal situation, phonetic research on every language would be assisted by forced alignment systems specifically built for the language; alignments are better for the trained language than for others. The work here is novel since we investigate the temporal accuracy of alignment using an aligner trained on a different source language. To date, we know of no such study that examines the temporal aspects of alignment using an aligner trained on a different source language. This may relate to the differing goals of those working in speech technology and those working in phonetics. The latter may be more interested in the details of aligner accuracy because precision in alignment is considered a necessity for the automatic extraction of acoustic data for phonetic research.

In the case of endangered and minority languages, conditions are frequently far from ideal. The initial stage of corpus transcription requires substantial linguist and native-speaker expertise and time. Segmentation of these corpora for the purposes of building a language-specific aligner often falls outside the purview of documentation projects and it may require expertise that the documentary linguist does not possess. Given these conditions, the process of corpus segmentation can be aided by forced aligners trained on other, more common languages.

There are a few factors that one must consider in selecting such an aligner. First, the work here has shown that an aligner trained with a larger set of allophonic phones can improve agreement for a language on which the aligner was not trained. Essentially, it is necessary to find an aligner with a phone set which reasonably closely matches the language's phonological system. Second, the data on which the aligner was trained can play an important role in the level of agreement. It is necessary to consider the nature of the corpus one wishes to segment prior to choosing the aligner. If the corpus consists of careful elicitations or word lists, an aligner based on read speech may do better than one based on spontaneous speech. If it consists of narratives and running dialogue, then an aligner based on spontaneous speech may be more appropriate. In general, alignment was poorer at utterance-initial and utterance-final

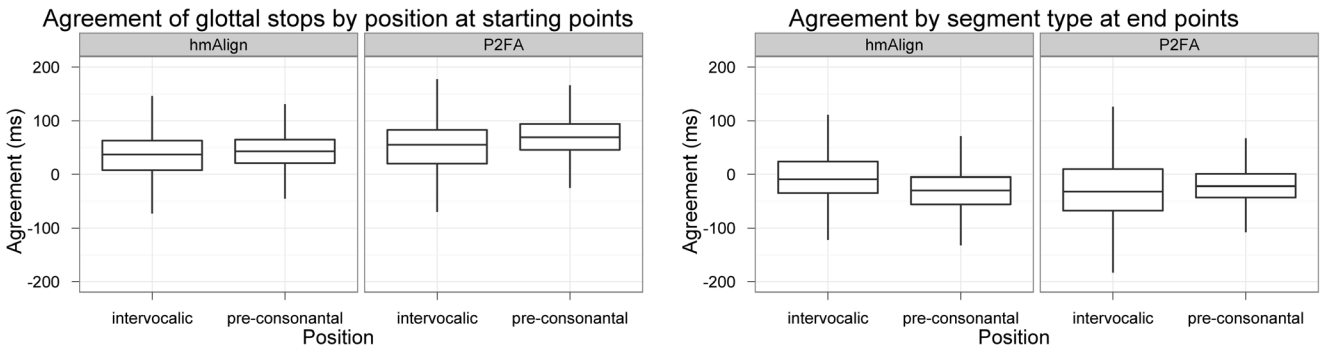

FIG. 6. Agreement for glottal stops across aligners.

positions for YM. This may be due to the language-specific nature of syllables in YM, which never have a coda. Languages with codas may match English-trained aligners better. This should be taken under consideration as well.

Once such factors have been considered, what alignment quality might one expect from forced alignment? In the data for HMALIGN, agreement was 71% at 30 ms and only 42% at 10 ms. While this may seem reasonably accurate for the YM data, the time needed to manually adjust erroneous data is a potential problem. In the current data, it might take longer to fix the remaining 29% (or 58%) than to simply hand-label all segments from scratch. Such a possibility is disheartening, but there is evidence that such adjustments might take less time than expected. First, not all natural classes are equally badly-aligned. In the current data with HMALIGN, fricatives and affricates were aligned at 99% at 20 ms, while stops and nasals were aligned at 80% at 20 ms. Among consonants, agreement was lowest for approximants. In the case of YM, these differences make more targeted manual adjustments possible. Adjustments to many obstruents can be ignored, while approximants must be more carefully examined. Moreover, since agreement for forced alignment is much lower at utterance boundaries, these positions can be given more attention than word-internal segmental boundaries. An alternative option is to use the initial alignment from an English aligner, such as HMALIGN, as the starting place for retraining. This would bring the phone models into better alignment with the acoustics and further reduce the amount of manual adjustment that is needed.

The importance of agreement accuracy may also vary depending on the motivation for the forced alignment. If the goal of alignment is an examination of non-local phonetic measures, like overall duration, formant structure, and mean F_0, then higher error rates are less consequential (assuming that error agreement follows a relatively normal distribution). If the goal is to examine phonetic measures for which very accurate alignment is essential, like stop VOT, formant transitions, and voicing-related F_0 perturbations, then substantial error in forced alignment is more problematic. While forced alignment with high agreement is important for phonetic data analysis, the presence of some error is less troublesome for other research goals.

TABLE XI. Agreement at start points with hand-labeling for glottal stops across aligners.

	P2FA		HMALIGN	
Threshold	V?V	V?CV	V?V	V?CV
10 ms	2.0%	2.7%	13.1%	9.5%
20 ms	5.7%	6.4%	25.7%	19.4%
30 ms	10.3%	11.5%	35.7%	30.8%
40 ms	15.7%	18.9%	46.3%	43.5%
50 ms	21.4%	28.2%	59.6%	57.4%

B. Comparison between aligners

In general, greater agreement was found for the YM data with HMALIGN than with P2FA. The main cause for the difference in agreement between aligners was assumed to be the phone set. The aligners had different phone sets. HMALIGN uses a phone set which is largely allophonic for English (context-sensitive). This includes variants for clear and velarized /l/, unaspirated and aspirated stops, syllabic sonorants, and the positional variants of /ɹ/. P2FA uses a context independent phone set which is rather similar to the phonemic inventory for English and does not contain consonantal allophones. Among the consonants, the largest difference in agreement between the aligners was for stops. As YM has only unaspirated stops, the unaspirated stop phone set from HMALIGN was a closer fit for the language than the more generic phone set in P2FA. Including these additional allophones in the phone set was likely responsible for the error reduction in HMALIGN.

A similar inference can be made regarding agreement for the YM glottal stop. The phone set of HMALIGN included the glottalized allophone of English /t/, [t͡ʔ], as a separate phone. While agreement was low for glottal stops for both aligners, there was a substantial improvement in agreement with intervocalic glottal stops with HMALIGN. This difference in agreement can only be related to this additional phone in HMALIGN. Notably, no differences in agreement between aligners occurred for pre-consonantal glottal stop.

A few smaller differences between aligners also emerged. For example, agreement for w vowels with a rising tone was significantly better with HMALIGN than with P2FA. This difference may be attributable to the training data. While the topic has not been explored in great detail, research on speech rate shows a rate reduction during read speech compared to spontaneous speech (Laan, 1997 on English, Hirose and Kawanami, 2002 on Japanese). The delay of an F_0 peak in the production of pitch accents (peak-

delay) is closely correlated with the duration of the syllable on which the accent is aligned (Silverman and Pierrehumbert, 1990). It is possible that forced alignment based on spontaneous speech training data, like P2FA, will contain a greater proportion of words where F_0 maxima occur on the following syllable, due to peak delay. By contrast, a forced aligner trained on read speech, like HMALIGN, might contain a smaller proportion of such words. For the YM data, F_0 maxima are aligned at the right edge of the vowel (DiCanio *et al.*, 2012). P2FA would have a disadvantage with the alignment of YM vowels with rising tones.

In addition, there may have been an effect of the match or mismatch of the type of utterances the aligners were trained on. P2FA is trained on the SCOTUS corpus, which contains speech from running court arguments (Yuan and Liberman, 2008). HMALIGN is trained on the TIMIT corpus, which contains only read speech. While SCOTUS arguments are not completely spontaneous (often being rehearsed), they are much less scripted than read speech. Given the training data, HMALIGN can be expected to be better suited to the forced alignment of words produced in isolation. This is indeed what happened with the YM corpus, which consists of words in isolation. Moreover, as P2FA is trained on running speech, one anticipates that it will be more sensitive to pauses than HMALIGN. Running speech is characterized by longer and more frequent pauses, more frequent hesitations, and shorter prosodic units (see Mehta and Cutler, 1988, and references therein). Here, out of 83 768 aligned boundaries, 5232 short pauses were inserted by P2FA. None were inserted by HMALIGN. These additional pauses were inserted during background noise in the recording or word-medially during words with silent intervals, e.g., during a long closure duration or during glottal closure with /ʔ/. Further tests of the two aligners on the same language but with spontaneous speech (more similar to SCOTUS) would be needed to test this possibility.

V. CONCLUSIONS

Comparisons between the P2FA and HMALIGN forced alignment systems show the latter system to provide better agreement than the former for corpus data from Yoloxóchitl Mixtec. This finding confirms the two hypotheses tested in the current study. First, forced alignment systems based on different languages can be successfully utilized for under-resourced languages, of which endangered languages are a particular case. Second, the inclusion of context-sensitive phones (HMALIGN) results in a 15.2% error reduction over a general set (P2FA), supporting the idea that context-sensitivity improves aligner performance (Sim and Li, 2008a, 2008b). In particular, agreement accuracy differences were most robust for stop consonants and for glottal stops. Moreover, these findings also argue that the validity of a forced aligner for the segmentation of a novel language depends closely on the style of discourse that is to be aligned. In general, data from endangered and minority language corpora have presented unique problems for automatic methods of segmentation. Given the paucity of training corpora and the difficulty in obtaining automatic aligners for such languages, forced alignment systems based on different languages offer a useful alternative or supplement to hand-labeled segmentation.

ACKNOWLEDGMENTS

The YM corpus was elicited by Castillo García, Amith, and DiCanio with support from Hans Rausing Endangered Language Programme Grant No. MDP0201 and NSF Grant No. 0966462. The authors would like to thank Leandro DiDomenico for his help with transcription labeling. This work was supported by NSF Grant No. 0966411 to Haskins Laboratories. The first two authors listed contributed equally to the current manuscript.

Adda-Decker, M., and Snoeren, N. D. (**2011**). "Quantifying temporal speech reduction in French using forced speech alignment," J. Phonetics **39**, 261–270.

Baayen, R. H. (**2008**). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* (Cambridge University Press, Cambridge, UK).

Badenhorst, J., van Heerden, C., Davel, M., and Barnard, E. (**2011**). "Collecting and evaluating speech recognition corpora for 11 South African languages," Lang. Res. Eval. **45**(3), 289–309.

Bates, D. M. (**2005**). "Fitting linear mixed models in R," R News **5**, 27–30.

Beddor, P. S., and Krakow, R. A. (**1999**). "Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation," J. Acoust. Soc. Am. **106**(5), 2868–2887.

Bickel, B., Banjade, G., Gaenszle, M., Lieven, E., Paudyal, N. P., Rai, I. P., Rai, M., Rai, N. K., and Stoll, S. (**2007**). "Free prefix ordering in Chintang," Language **83**(1), 43–73.

Boersma, P., and Weenink, D. (**2012**). "Praat: Doing phonetics by computer" [computer program], www.praat.org (date last viewed 10/1/12).

Boula de Mareüil, P., Corredor-Ardoy, C., and Adda-Decker, M. (**1999**). "Multi-lingual automatic phoneme clustering," in *Proceedings of the 14th International Congress of the Phonetic Sciences*, San Francisco, CA, pp. 1209–1212.

Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Povey, D., Rastrow, A., Rose, R. C., and Thomas, S. (**2010**). "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4334–4337.

Dalsgaard, P., Andersen, O., and Barry, W. (**1991**). "Multi-lingual label alignment using acoustic- phonetic features derived by neural-network technique," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, 197–200.

Dankovičová, J. (**1997**). "Czech," J. Int. Phonetic Assoc. **27**(1), 77–80.

DiCanio, C., Amith, J., and Castillo-García, R. (**2012**). "Phonetic alignment in Yoloxóchitl Mixtec tone," talk presented at *The Society for the Study of the Indigenous Languages of the Americas, Annual Meeting*, Portland, OR.

Du Bois, J. W. (**1987**). "The discourse basis of ergativity," Language **63**(4), 805–855.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (**1993**). *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Linguistic Data Consortium, Philadelphia).

Hai, D. V., Xiao, X., Chng, E. S., and Li, H. (**2012**). "Context dependent phone mapping for cross-lingual acoustic modeling," in *Proceedings of ISCSLP*, pp. 16–20.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (**2012**). "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," IEEE Signal Process. Mag. **29**(6), 82–97.

Hirose, K., and Kawanami, H. (**2002**). "Temporal rate change of dialogue speech in prosodic units as compared to read speech," Speech Commun. **36**, 97–111.

Hosom, J-P. (**2009**). "Speaker-independent phoneme alignment using transition-dependent states," Speech Commun. **51**, 352–368.

Huffman, M. K. (**2005**). "Segmental and prosodic effects on coda glottalization," J. Phonetics **33**, 335–362.

Hura, S. L., Lindblom, B., and Diehl, R. L. (**1992**). "On the role of perception in shaping phonological assimilation rules," Lang. Speech **35**(1–2), 59–72.

Imseng, D., Bourlard, H., and Garner, P. N. (**2012**). "Boosting under-resourced speech recognizers by exploiting out of language data—A case study on Afrikaans," in *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages*, pp. 60–67.

Jarifi, S., Pastor, D., and Rosec, O. (**2008**). "A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis," Speech Commun. **50**, 67–80.

Jones, D., and Ward, D. (**1969**). *The Phonetics of Russian* (Cambridge University Press, Cambridge, UK).

Karafiát, M., Janda, M., Černocky, J., and Burget, L. (**2012**). "Region dependent linear transforms in multilingual speech recognition," in *Proceedings from ICASSP 2012*, pp. 4885–4888.

Krauss, M. (**1992**). "The world's languages in crisis," Language **68**, 4–10.

Laan, G. P. M. (**1997**). "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," Speech Commun. **22**, 43–65.

Lei, X., Hwang, M-Y., and Ostendorf, M. (**2005**). "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR," in *Proceedings of Interspeech–2005*, Lisbon, Portugal, pp. 2981–2984.

Lin, C-Y., Roger Jang, J-S., Chen, K-T. (**2005**). "Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS," Comput. Ling. Chinese Lang. Process. **10**(2), 145–166.

Livescu, K., Fosler-Lussier, E., and Metze, F. (**2012**). "Subword modeling for automatic speech recognition: Past, present, and emerging approaches," IEEE Signal Process. Mag. **November**, 44–57.

Macken, M. A., and Salmons, J. C. (**1997**). "Prosodic templates in sound change," Diachronica **14**(1), 31–66.

Malfrère, F., Deroo, O., Dutoit, T., and Ris, C. (**2003**). "Phonetic alignment: Speech synthesis-based vs. Viterbi-based," Speech Commun. **40**, 503–515.

Manuel, S. Y. (**1991**). "Some phonetic bases for the relative malleability of syllable-final versus syllable-initial consonants," in *Proceedings of the 12th International Congress of Phonetic Sciences*, Université de Provence, Aix-en-Provence, Vol. 5, pp. 118–121.

Mehta, G., and Cutler, A. (**1988**). "Detection of target phonemes in spontaneous and read speech," Lang. Speech **31**(2), 135–156.

Ní Chasaide, A., Wogan, J., Ó Raghallaigh, B., Ní Bhriain, Á., Zoerner, E., Berthelsen, H., and Gobl, C. (**2006**). *Speech Technology for Minority Languages: The Case of Irish (Gaelic), in INTERSPEECH–2006*, pp. 181–184.

Ohala, J. (**1990**). "The phonetics and phonology of aspects of assimilation," Papers Lab. Phonol. **1**, 258–275

R Development Core Team (**2012**). "R: A language and environment for statistical computing" [computer program], http://www.R-project.org, R Foundation for Statistical Computing, Vienna, Austria (date last viewed 10/1/12).

Rabiner, Lawrence, R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition, Prentice-Hall Signal Processing Series* (Prentice Hall, Englewood Cliffs, NJ).

Roux, J. C., and Visagie, A. S. (**2007**). "Data-driven approach to rapid prototyping Xhosa speech synthesis," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pp. 143–147.

Silverman, K., and Pierrehumbert, J. (**1990**). "The timing of prenuclear high accents in English," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. Beckman, Cascadilla Proceedings Project, Somerville, MA, pp. 103–112.

Sim, K. C., and Li, H. (**2008a**). "Robust phone mapping using decision tree clustering for cross-lingual phone recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4309–4312.

Sim, K. C., and Li, H. (**2008b**). "Context-sensitive probabilistic phone mapping model for cross-lingual speech recognition," In *Proceedings of Interspeech 2008*, International Speech Communication Association (ISCA), pp. 2715–2718.

Whalen, D. H., and Simons, G. F. (**2012**). "Endangered language families," Language **88**, 155–173.

Whalen, D. H., and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica **49**, 25–47.

Yarrington, D., Pennington, C., Bunnell, H. T., Gray, J., Lilley, J., Nagao, K., and Polikoff, J. (**2008**). "ModelTalker voice recorder (MTVR)—A system for capturing individual voices for synthetic speech," talk presented at the *ISAAC 13th Biennial Conference*, Montreal, Canada (August 2–7).

Yip, M. (**2002**). *Tone*, *Cambridge Textbooks in Linguistics* (Cambridge University Press, Cambridge, UK), p. 376.

Yuan, J., and Liberman, M. (**2008**). "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics 2008*, pp. 5687–5690.

Yuan, J., and Liberman, M. (**2009**). "Investigating /l/ variation in English through forced alignment," in *INTERSPEECH–2009*, pp. 2215–2218.

Zue, V. W., and Seneff, S. (**1996**). "Transcription and alignment of the TIMIT database," in *Recent Research towards Advanced Man-Machine Interface through Spoken Language*, edited by H. Fujisaki (Elsevier, Amsterdam), pp. 515–525.

Zue, V., Seneff, S., and Glass, J. (**1990**). "Speech database development at MIT: TIMIT and beyond," Speech Commun. **9**, 351–356.