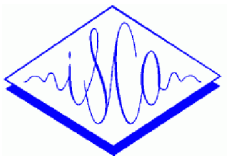

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Assessing agreement level between forced alignment models with data from endangered language documentation corpora

Christian T. DiCanio[1], *Hosung Nam*[1], *D. H. Whalen*[1,2,3], *H. Timothy Bunnell*[4,5], *Jonathan D. Amith*[6,7], *Rey Castillo García*[8]

[1]Haskins Laboratories, New Haven, CT, USA.
[2]Speech-Language-Hearing Program, CUNY Graduate Center, New York, NY, USA.
[3]Endangered Language Fund, New Haven, CT, USA.
[4]Nemours Biomedical Research, Alfred I. duPont Hospital for Children, Wilmington, DE, USA.
[5]Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA.
[6]Department of Anthropology, Gettysburg College, Gettysburg, Pennsylvania, USA.
[7]Smithsonian Institution, Washington D.C., USA.
[8]CIESAS, Mexico City, D.F., Mexico.

Christian T. DiCanio dicanio@haskins.yale.edu, Hosung Nam nam@haskins.yale.edu,
Douglas H. Whalen whalen@haskins.yale.edu, H. Timothy Bunnell bunnell@asel.udel.edu
Jonathan D. Amith jonamith@gmail.com, Rey Castillo García castagr@hotmail.com

10.21437/Interspeech.2012-43

Abstract

Automatic forced alignment between transcriptions has achieved high levels of agreement for languages with large corpora, but the technique holds great promise for work on all languages. Here, we apply two forced alignment programs to data from an endangered Mixtecan language of Mexico. Both yielded a majority of boundaries within 20 ms of hand-labeled ones. Phonemes with fairly steady-state elements (e.g. nasals, fricatives) were more accurately labeled than others. Forced alignment thus may increase efficiency of labeling texts from smaller languages, at least in cases where the phoneme inventories are similar to those of the languages of the training.

Index Terms: speech recognition, phonetics, linguistics

1. Introduction

Over the past twenty years, the documentation of endangered languages has become a vital topic in the study of linguistics. Part of what motivates this interest is the need to provide a record of linguistic and cultural diversity that is on the brink of extinction. Dozens of projects throughout the world are currently devoted to the collection of language materials in endangered or minority languages and to the description of these languages' phonological and grammatical structures. The collection of speech corpora for these languages, including texts and conversations, comprises most of the data for descriptive and exploratory linguistic analyses. However, the utility of these corpora for quantitative phonological and phonetic analyses is not well-established. One of the reasons for this gap is the excessive time demands required for manual speech annotation. Once proper phonemic transcription of speech corpora is completed by linguists familiar with the language's structure, these transcriptions must be acoustically-aligned with the recording itself (labeling) if further computational analysis is to be possible. This labeling process is even more time-consuming than transcription, yet only once this is done are researchers able to use this data to analyze naturally-occurring speech samples in these languages.

One way to accelerate this labeling process is by using forced alignment, which is performed by an HMM-based speech recognizer. Forced alignment can derive word- and phone-level labeling from a word-level transcription and a dictionary. Hence, forced alignment has been useful for automatic labeling in well-studied languages such as English, French, and Mandarin Chinese [1, 2, 3] and in less well-documented languages like Gaelic, the official language of Ireland, and Xhosa, an official language of South Africa [4, 5]. In developing a speech recognizer for a new language, forced alignment from a different language with a well-trained recognizer can be used to provide a preliminary phone segmentation. In this study, we compare the accuracy of two English-trained forced aligners: P2FA (the Penn Phonetics Lab Forced Aligner) [6] and hmAlign [7] on a controlled corpus of Yoloxóchitl Mixtec (YM; ISO 639 code xty) speech. hmAlign is a stand-alone version of the aligner developed for the ModelTalker TTS system voice recording program [7] (see §6). YM is an endangered language spoken in Guerrero, Mexico and the focus of a large-scale language documentation project (see §6). The analysis of forced alignment results using data from this language provides an assessment of its utility for endangered language corpora and a comparison between two different forced aligners.

1.1. Motivations

A common procedure for using forced alignment on minority languages involves the initial extraction of alignments using a model trained on a baseline HMM-trained forced alignment system. However, there has been no systematic analysis of the success of this step in creating language-specific ASR systems. One reason for this is the general difficulty in making a direct comparison between forced alignment systems. Minor differences in implementation prevent a clear benchmark comparison between models [8]. By presenting the agreement results from two untrained forced aligners using an identical corpus, it is possible to make such a direct comparison. Furthermore, we faced certain challenges using forced alignment with this corpus and would like to report how different aligners perform in relation to it.

The choice of using the YM language was based on both the availability of good corpus data and the structure of the language's phonological system. Data from many languages has become readily available from large-scale documentation

projects [9]. YM is a useful language because there is a large amount of accurately transcribed corpus data. The language was also ideal for analysis because despite its complex tonal system, its segmental inventory is relatively simple. As the forced aligners were each trained on English, the English phoneme set could be used for transcribing Mixtec segments.

2. Background

2.1. Language Background & Corpus

The phonological inventory of YM is smaller than that of English, but with a few important differences. First, there is one series of stops, /p, t, k, k^w/, all of which are voiceless and unaspirated. Within this set, the labialized velar stop is a distinct phoneme while it is a sequence of two phonemes in English. This was coded as /k/ in the forced alignment. Second, there are two prenasalized stops, /mb/ and /nd/. These sounds only occur as sequences in English, e.g. *wi**nd**ow,* and are not phonemic units. They were coded as simple nasals. Third, glottal stops are phonemic in YM and surface both word-medially in pre-consonantal position before sonorants, e.g. /sa$ʔ^3$ma^4/ *'napkin',* and in intervocalic position, e.g. /ndo^{1}ʔo^4/ *'basket.'* The entire consonant inventory is shown in Table 1. Segments in parentheses are rare both in the language and in the corpus.

The vowel inventory is a typologically common five vowel system: /i, e, a, o, u/. Each vowel may also be contrastively nasalized: /ĩ, ẽ, ã, õ, ũ/. Vowels are never reduced in YM words, but always surface as full variants. All vowels following nasal consonants undergo obligatory progressive nasalization, e.g. /nũ^{14}ũ^3/ *'face'*. Nasalization only occurs contrastively on word-final syllables [10]. Note that nasality was not marked within vowel segments, so that treating nasalized vowels as equivalent to English oral vowels yielded the same number of boundaries.

Content words in YM are minimally bimoraic and maximally trimoraic. Minimal words are either disyllabic, with two short vowels, e.g. CVCV, or monosyllabic with a long vowel, e.g. CVV. Maximal words are either disyllabic, with a short and long vowel, e.g. CVCVV, or trisyllabic with three short vowels, e.g. CVCVCV. Long vowels only occur in word-final syllables.

The language has a complex tonal system, where each tone is indicated with numerals following each vowel. There are 10 tonal patterns which contrast on the final mora of a word and 5 on the penultimate mora. All syllables are open [10].

	Bilabial	Dental	Post-alveolar	Palatal	Velar	Labialized Velar	Glottal
Plosive	(p)	t			k	k^w	ʔ
Pre-nasalized plosive	(mb)	nd					
Affricate			tʃ				
Nasal	m	n					
Tap		(ɾ)					
Fricative	β	s	ʃ				
Approximant		l		j			

Table 1: *Consonant inventory of Yoloxóchitl Mixtec.*

2.2. Corpus

The corpus data consisted of 2 hours of YM speech, containing 261 words repeated 6 times by 5 different native speakers. While a few words were produced with tonal changes that indicated negation or aspect marking, no overt segmental prefixes appeared on any word. A total of 7830 word repetitions were examined, comprising 27166 speech segments. All words were produced in citation form. Hand-labeling was done by the first author and compared to the forced alignments done by the P2FA and hmAlign models. Another set of labels provided both a phonemic transcription and encoded the location of the consonant or vowel in the word, e.g., C1 for the first consonant, V2 for the second vowel. This labeling allowed for the analysis of error reduction assessment by position in the word.

2.3. Forced alignment models

P2FA's acoustic models are GMM-based monophone-HMMs trained using the SCOTUS corpus, which includes oral arguments from the Supreme Court of the United States. Each HMM state consists of 32 Gaussians Mixture components on 39 Perceptual Linear Predictive (PLP) coefficients (12 Cepstra plus energy, delta and acceleration). P2FA employs CMU phones, which do not show allophonic variants in English. On the other hand, hmAlign uses a set of discrete monophone HMMs trained on data from the TIMIT corpus. For hmAlign, separate 375-word codebooks were trained on vectors of 8 Mel Frequency Cepstral Coefficients (MFCC), plus their delta and acceleration coefficients. Unlike P2FA, the list of phones used in hmAlign includes some allophonic variants. Aspirated and unaspirated stops are distinguished, as is the glottalized coda [t] variant. YM contains unaspirated stops and a contrastive glottal stop. The phone-phoneme mappings are shown in Table 2.

Two steps were taken before running the aligners. First, the Mixtec speech data were downsampled from 48kHz to 16kHz for both P2FA and hmAlign. Second, in each model, a pronunciation dictionary was constructed where the pronunciations of the YM word were coded. Table 2 shows the phonemic inventory of YM and the phonetic exponents in the language. The symbol used in each of the forced aligners is also given along with its phonetic exponents.

Mixtec	P2FA	hmAlign
/p/ [p]	P [p^h, p]	PP [p]
/t/ [t]	T [t^h, t, t͡ʔ, 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, t͡ʔ, ɾ]	TQ [t͡ʔ]
/nd/ [nd]	N [n]	NN [n]
/tʃ/ [tʃ]	CH [tʃ]	CH [tʃ]
/m/ [m]	M [m]	MM [m]
/n/ [n]	N [n]	NN [n]
/β/ [β, w, 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]	II [i]
/e/ [e, ɛ]	EH [ɛ]	EH [ɛ]
/a/ [a]	AA [ɑ]	AA [ɑ]
/o/ [o, ɔ]	AO [ɔ]	AO [ɔ]
/u/ [u]	UW [u, ʉ]	UW [u, ʉ]

Table 2: *Mixtec phonemes with corresponding P2FA and hmAlign phones.*

3. Analysis

3.1. General Results

The hmAlign system performed better overall on the corpus than the P2FA system. Overall, agreement of hmAlign is 62.9% within 20 ms, compared with 52.7% within 20 ms for the P2FA system. These results for hmAlign reflect a 21.6% relative reduction in error over P2FA. Performance results within 10-40 ms are given in Table 3.

Threshold	hmAlign	P2FA
10 ms	45.3%	35.8%
20 ms	62.9%	52.7%
30 ms	71.6%	65.2%
40 ms	80.9%	73.6%

Table 3: *Agreement with manual labeling*

Within different ranges, hmAlign retains a significant reduction in error. Since neither P2FA nor hmAlign were trained on Mixtec data, agreement levels are low in comparison with trained models [8]. Furthermore, the corpus is entirely composed of words produced in isolation. It is known that utterance boundaries increase errors in automatic speech recognition [11]. Since the initial consonant and final vowel for each word are coincidental with utterance boundaries, one anticipates greater errors to be produced at word edges than word-internally. For both hmAlign and P2FA, word-final vowels were aligned less accurately than word-internal vowels. However, greater agreement was found for utterance-initial consonants than for utterance-internal consonants. These data are shown in Figure 1. In this figure, "c1" is the word and utterance-initial consonant in disyllabic words, while "c2" is the word-medial consonant. The label "v1" is the word-medial vowel, while "v2" is the word and utterance-final vowel.

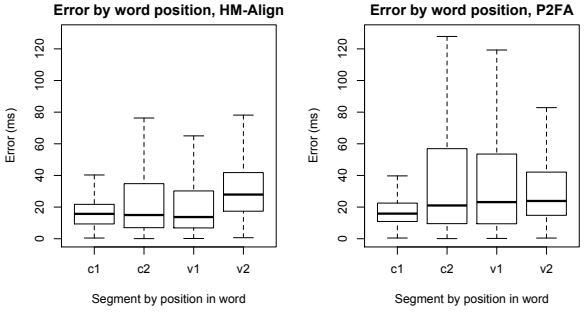

Figure 1: *Alignment errors by segment position. Whisker boundaries reflect alignment error maxima and minima, box boundaries reflect the upper and lower quartiles, and the solid line reflects the median value. Outliers are removed for visualization purposes.*

3.2. Phoneme-specific and class-specific differences

The hmAlign system performs better than the P2FA system across most sound types. Notably though, there is a greater increase in model accuracy for glottal segments. In general, one predicts low agreement for glottal stops surfacing in intervocalic position, as they do not occur in this environment in English. However, there is a similarity between glottal stops surfacing in pre-consonantal position in Mixtec, e.g. /ʃaʔ^{4}ni^{24}/, *'kills',* and glottalized variants of /t/ in English, e.g. /tʃʌt͡ʔni/, *'chutney.'* Since the hmAlign system encodes this glottalized /t/ variant explicitly as a phone, one predicts greater error reduction in the hmAlign system for these segments than for the more unnatural /VʔV/ sequences.

However, this is not observed. There is greater agreement for intervocalic glottal stop than the preconsonantal one in the hmAlign system; it is the reverse in P2FA. These data are shown in Table 4. Data showing the overall accuracy of /ʔ/ detection are shown in Figure 2.

	P2FA		hmAlign	
Accuracy	VʔV	VʔCV	VʔV	VʔCV
10 ms	3.9%	12.1%	15.4%	13.7%
20 ms	8.9%	23.3%	29.9%	25.4%
30 ms	16.1%	34.9%	40.9%	39.0%

Table 4: *Agreement between models by position of glottal stop.*

While agreement is quite low for /ʔ/ detection in both forced aligners, the relative reduction in error for the hmAlign system for intervocalic /ʔ/ is high, at 30% within 30 ms.

Examining agreement across natural classes, consonants composed primarily of steady state acoustic portions (fricatives, nasals) had higher agreement in both forced aligners than consonants with transitions (vowels, approximants, stops, affricates). These data are shown in Figure 3. As a result, increases in model agreement for fricatives and nasals resulted in greater error reduction between models. With respect to oral stops, the agreement for the hmAlign system is 53.9% within 20 ms, but 46.0% within 20 ms for P2FA. This reflects a 14.6% error reduction between models. For fricatives, the agreement for hmAlign is 86.8% within 20 ms and 81.1% within 20 ms for P2FA. This reflects a 30.2% error reduction in the hmAlign system over P2FA.

For nasal consonants, agreement performance is 73.0% within 20 ms for P2FA and 78.6% within 20 ms for hmAlign. This reflects a 20.7% error reduction. Model differences are smaller for approximant and affricate tokens. Agreement for approximants is 44.3% within 20 ms for P2FA and 50.1% within 20 ms for hmAlign. Agreement for affricates is 55.6% within 20 ms for P2FA and 61.7% within 20 ms for hmAlign.

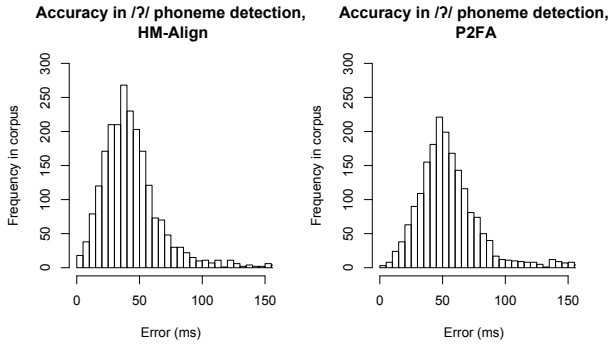

Figure 2: *Accuracy in glottal stop detection across models.*

3.3. Issues in assessing agreement

The observed differences between the forced aligners are robust. However, the estimation of agreement relies on an averaging of both manual-machine alignment for the onset transition of a phone and the manual-machine alignment for its offset. This method assumes that agreement is similar for onsets and offsets, but it is not. In general, onset transitions at the left edge of consonants, e.g. V-C, show lower agreement than offset transitions at the right edge of consonants, e.g. C-V. Agreement at onset transitions is 39.8% within 20 ms for

P2FA and 48.6% within 20 ms for hmAlign. In contrast, agreement at offset transitions is 70.8% within 20 ms for P2FA and 79.1% within 20 ms for hmAlign. The differences are even more robust with respect to certain sound types. For instance, agreement on stop onset transitions in the hmAlign model are just 23.0% within 20 ms, but 84.8% within 20 ms on stop offset transitions. This asymmetry in alignment agreement mirrors differences in the human perception of C-V and V-C transitions, where acoustic cues are often weaker in V-C transitions [12, 13].

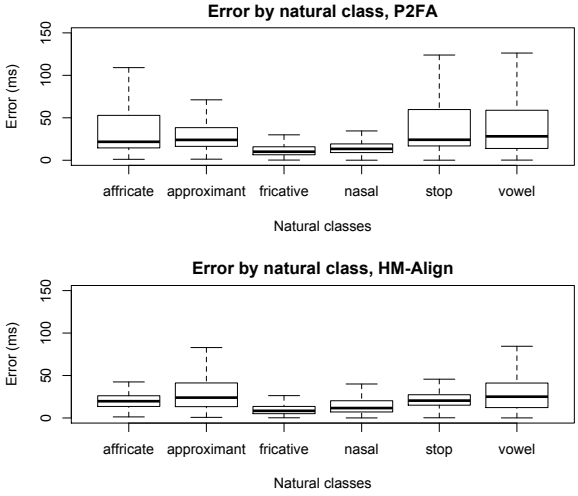

Figure 3: Error by segment natural class.

4. Discussion

Overall, the hmAlign system performed better on the corpus than P2FA. Part of the increased error reduction in hmAlign is a product of varying phone sets for transcription. While P2FA uses a mostly phonemic level English transcription, hmAlign contains allophonic English variants. Some of these variants more closely match the phonetic variants produced in the YM corpus. In particular, a substantial increase in error reduction was found for glottal stop alignment in the hmAlign system. While glottal stops are not phonemic in English, glottalization occurs concomitantly with tongue tip constriction in the production of English coda [t] [14]. Since the forced alignment system in hmAlign used a phone built explicitly with glottalization, it was more accurate aligning glottal stops than P2FA, which did not contain this phone. Moreover, the presence of an unaspirated stop phone in the hmAlign system more closely matched the phonetic structure of YM stops. This resulted in substantial error reduction in the hmAlign model.

Each of the current models that we examined has low agreement performance with respect to hand-labeling. It is hoped that these initial alignments will be useful for bootstrapping a language-specific model. However, while language-specific models show increased agreement, we emphasize that their utility is contingent on the availability of large sets of transcribed corpora which have already been labeled. One of the major bottlenecks in descriptive phonetic research of endangered languages is manual phonetic labeling. Language-specific forced aligners are often built on large corpora on which manual phonetic labeling has already been completed. For many endangered languages, such corpora are not available. In these cases, researchers may be able to use existing forced alignment systems which show reasonably good agreement with hand-labeling. Assessing the viability of using untrained models aids this goal.

5. Conclusions

Forced alignment for an untrained language yielded useful results even though no retraining of the aligners was performed. The language we chose had a segment inventory that was similar to a subset of the English inventory that the aligners were trained on. In this context, it would appear that using a forced alignment as a first pass for labeling the speech signal would save time. It is conceivable that retraining existing aligners on a new language with small amounts of data would improve performance further, allowing an even wider usage of this efficient technique.

6. Notes/Acknowledgements

For questions regarding the availability of hmAlign, please contact the fourth author. The YM corpus was elicited by Castillo García, Amith, and DiCanio with support from Hans Rausing Endangered Language Programme Grant MDP0201 and NSF grant 0966462. The authors would like to thank Leandro DiDomenico for his help with transcription labeling. This work was supported by NSF grant 0966411 to Haskins Laboratories.

7. References

[1] Adda-Decker, M., Snoeren, N. D. "Quantifying temporal speech reduction in French using forced speech alignment", Journal of Phonetics, 39:261-270, 2011.

[2] Lin, C-Y., Roger Jang, J-S., Chen, K-T. “Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpora for Concatenation-based TTS”, Computational Linguistics and Chinese Language Processing, 10(2):145-166, 2005.

[3] Yuan, J., Liberman, M. "Investigating /l/ Variation in English Through Forced Alignment", in INTERSPEECH-2009, 2215-2218, 2009.

[4] Ní Chasaide, A., Wogan, J., Ó Raghallaigh, B., Ní Bhriain, Á., Zoerner, E., Berthelsen, H., Gobl, C. “Speech Technology for Minority Languages: the Case of Irish (Gaelic)”, INTERSPEECH-2006, 181-184, 2006.

[5] Roux, J.C., Visagie, A.S. “Data-driven Approach to Rapid Prototyping Xhosa Speech Synthesis”, in Proc. of the 6th ISCA Workshop on Speech Synthesis, 143-147. 2007.

[6] Yuan, J., Liberman, A. M., "Speaker Identification on the Scotus Corpus", in Proceedings of ICASSP-2008. 2008.

[7] Bunnell, H.T., Pennington, C., Yarrington, D., Gray, J. "Automatic Personal Synthetic Voice Construction", INTERSPEECH-2005. 89-92. 2005.

[8] Hosom, J-P. "Speaker-independent phoneme alignment using transition-dependent states", Speech Communication, 51:352-368. 2009.

[9] Simons, G., & Bird, S. "The Open Language Archives Community: An infrastructure for distributed archiving of language resources", Literary and Linguistic Computing, 18: 117-128. 2003.

[10] Castillo García, R. Descripción fonológica, segmental, y tonal del Mixteco de Yoloxóchitl, Guerrero. M.A. Thesis. CIESAS, Mexico, D.F. 2007.

[11] Goldwater, S., Jurafsky, D., Manning, C.D. "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that decrease speech recognition error rates", Speech Communication, 52:181-200. 2010.

[12] Ohala, J. "The phonetics and phonology of aspects of assimilation", in Papers in Laboratory Phonology, 1:258-275. 1990.

[13] Hura, S.L., Lindblom, B., Diehl, R.L. "On the role of perception in shaping phonological assimilation rules", Language and Speech, 35(1-2):59-72. 1992.

[14] Huffman, M. "Segmental and Prosodic effects on coda glottalization", Journal of Phonetics, 33:335-362. 2005.