

Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus

Lei Chen¹, Yang Liu^{1,2}, Mary Harper¹, Eduardo Maia¹, Susan McRoy³

¹ Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907

{chenl,yangl,harper,maia}@ecn.purdue.edu

² The International Computer Science Institute
Berkeley, CA 94704

³ Computer Science
University of Wisconsin-Milwaukee
Milwaukee, WI 53211

Abstract

People, when processing human-to-human communication, utilize everything they can in order to understand that communication, including speech and information such as the time and location of an interlocutor's gesture and gaze. Speech and gesture are known to exhibit a synchronous relationship in human communication; however, the precise nature of that relationship requires further investigation. The construction of computer models of multimodal human communication would be enabled by the availability of multimodal communication corpus annotated with synchronized gesture and speech features. To investigate the temporal relationships of these knowledge sources, we have collected and are annotating several multimodal corpora with time-aligned features. Forced alignment between a speech file and its transcription is a crucial part of multimodal corpus production. This paper investigates a number of factors that may contribute to highly accurate forced alignments to support the rapid production of these multimodal corpora including the acoustic model, the match between the speech used for training the system and that to be force aligned, the amount of data used to train the ASR system, the availability of speaker adaptation, and the duration of the alignment segment.

1. Introduction

Natural human communication is a rich interaction involving a variety of auditory and visual channels (Quek et al., 2002). To investigate the relationships among these channels in dialog (e.g., words, prosody, head position, gaze, and gesture), we have collected and are annotating several multimodal corpora with time-aligned features from several channels. Human dialogs were videotaped using a set of cameras that were both temporally and spatially calibrated (see <http://vislab.cs.wright.edu/KDI/>). The audio was digitally recorded using a unidirectional boom mounted microphone placed at a fixed distance for each interlocutor. These dialogs were recorded in a somewhat noisy laboratory environment and contain much speech overlap. Gestures and other visual features of each interlocutor are tracked algorithmically from the video, and the speech is transcribed and time-aligned. The elements of the corpus, such as word boundaries, 3D hand position, and other gesture features, such as effort and holds, are being temporally aligned to support measurement studies and data-driven modeling of multimodal communication.

With the availability of synchronized multimodal corpora, relationships among the different channels can be studied. Our initial efforts have focused on the relationship between gesture and spoken words within this framework. There are two reasons for selecting word-level, rather than phoneme-level granularity in the audio data. First, gesture tends to be super-segmental and so it is more likely to be related to longer stretches of speech (i.e., syllables, words, and phrases) than with the finer structure of words. Second, since the video sampling rate is much lower than for speech (29.97 Hz in the NTSC format compared to 44.1 KHz downsampled to 16 KHz), alignment to phonemes would be difficult to achieve with the technology used in collecting the corpus. Therefore, we focus on the time-alignment of audio segments at the syllable level and above in our investigations. For example, given a corpus of temporally aligned words, speech repair components (e.g., the reparandum, editing phrase, and alteration), and gesture patterns, we were able to develop a deeper understanding of how gestures are used to mark speech repairs (Chen et al.,

2002).

Word alignments can be obtained either by manually or automatic alignment. Manual alignment is time and resource consuming, although potentially more accurate than a fully automatic method. Hence, a semi-automatic procedure (Kessens and Strik, 2003) is widely used. This typically involves an initial forced alignment of a speech transcription to its audio file using an HMM-based automatic speech recognition (ASR) system, followed by a manual human check-and-repair step. For large scale corpus production, reducing human effort is of particular concern. The quality of the initial forced alignment directly impacts the amount of human effort required in the second step. If fewer and smaller errors are made, then less human effort is required in the second step.

There have been surprisingly few studies investigating the impact of the ASR system on forced alignment accuracy. Many times public availability of a system is the deciding factor for selecting it for forced alignments (e.g., the Aligner tool (Wightman and Talkin, 1997) was used to prepare the Variation in Conversation (ViC) corpus (Pitt et al., 2003), and the ISIP ASR system with a mono-phone acoustic model (R.Sundaram et al., 2000)) was used to produce the German Broadcast News corpus (Eickeler et al., 2002)). Although ASR systems have often been compared on word error rate and phonetic alignment accuracy, few studies have considered word alignment accuracy. Kessens (Kessens and Strik, 2003) investigated the effect of varying several properties of an ASR system on phonetic transcription accuracy and found that lower word error rates do not necessarily result in more accurate phonetic transcription. The question is, does this also hold for word forced alignments? Speech recognition systems are constructed to minimize word error rate, not to determine the most accurate time alignment of words to speech. Although it is likely that more accurate ASR systems would produce more accurate word forced alignments, this is currently an open question, especially in light of Kessens' findings.

Accurate phonetic alignment is important for acoustic model training, for constructing an acoustic inventory for speech synthesis, and for various speech science measure-

ment studies. Hence, research focusing on phonetic alignment accuracy has received some attention. See Hosom's Ph.D. thesis (Hosom, 2000) for a comprehensive review on phonetic alignment. Because the aim of word alignment is to obtain accurate word boundaries, the phonetic distinctions across words has greater importance than the phonetic distinctions within a word. Hence, there is also a need to investigate the factors of an ASR system that contribute to accurate word alignment.

There are several factors that may affect the quality of forced alignments that we will investigate in this paper. One set of factors is related to the ASR system: its model, the match between the speech used to train the ASR system and that to be force aligned, the amount of speech used for training, and the availability of speaker adaptation. Another factor concerns the duration of speech segments to be aligned. Our goal is to develop a better understanding of the factors that contribute to more accurate forced alignments in order to reduce the amount of tedious manual fixing, as we scale up multimodal corpus production.

2. Forced Alignment Experiments

For word forced alignments, four major components of the ASR system are used: the signal processing front-end, the trained HMM acoustic model, the lexicon, and the Viterbi decoding module. Across ASR systems, the signal processing front-ends are quite similar, and the lexicon is usually tailored to the specific application prior to forced alignment. Although many decoding passes are often used in modern speech recognition systems, only one decoding pass is typically used for forced alignment. Among the four components used for forced alignment, the acoustic models have the greatest variability across ASR systems and are likely to greatly impact forced alignment accuracy. Some factors related to the acoustic model include its modeling capability, the match between the speech used for training and that to be force aligned, and the amount of data used to train the ASR system. We will evaluate the impact of these factors by measuring the forced alignment accuracy of four different systems:

1. **Aligner.** Wightman (Wightman and Talkin, 1997) constructed this forced alignment tool using HTK to produce both word and phonetic alignments. Aligner's acoustic model, a mono-phone HMM model with five Gaussian mixtures, is trained on seven hours of read speech from the TIMIT corpus. As a publically available tool, Aligner is commonly used by many researchers to perform forced alignments.
2. **WSJ-HTK.** To investigate the impact of a more complex HMM topology and increased training set size holding read speech constant, we evaluate alignments using an HTK-based gender independent triphone HMM with eight Gaussian mixtures (16 for silence and short pause models) that uses a 39 cepstral coefficients. The model was trained on 57 hours of read speech from the Wall Street Journal ASR corpus.
3. **SWB-ISIP.** This system is built using the public domain ASR toolkit developed at Mississippi State. The acoustic model used is a context dependent model with 16 Gaussian mixtures trained on 60 hours of SwitchBoard 1 (2998 conversation sides) and 20 hours of English CallHome conversational speech for the 2000 Hub5 evaluation.
4. **SWB-SRI.** This system, built using SRI DECIPHER (Stolcke et al., 2000), uses gender-dependent context dependent HMM models with 64 Gaussian mixtures in addition to 39 cepstral coefficients, Vocal Tract Length Normalization, and cepstral mean and variance normalization are used for each speaker. The acoustic model was developed for the 2002 Hub5 evaluation and was trained on 160 hours of SwitchBoard (3094 conversations), 16 hours of English

CallHome conversational speech (100 conversations), and 18 hours of Macrophone read speech.

5. **BN-SRI.** This system is like SWB-SRI except that it was trained on 200 hours of speech from the Broadcast News ASR corpus, which is made up of largely teleprompted read speech.

The ASR systems above all use speaker independent (SI) models, which can be less accurate than an adequately trained speaker dependent (SD) acoustic models (Anastasakos et al., 1996). Although it is infeasible to construct a SD acoustic model for each speaker in a large corpus, it is possible to improve the accuracy of forced alignments by using speaker adaptation, wherein a small amount of data from a specific speaker is used to adjust the SI models in order to better represent that speaker. As in embedded training used for ASR acoustic model refinement, we will adapt our models using the speech to be aligned and its transcription. This avoids the increased costs associated with collecting and transcribing extra adaptation data. For this investigation, the method chosen for offline supervised adaptation was maximum likelihood linear regression (MLLR), which is available in HTK and SRI Decipher. Hence, we evaluate the impact of adaptation on the WSJ-HTK, BN-SRI, and SWB-SRI systems.

Forced alignment of an entire speech file to its transcription has the advantage that it is quite straight forward. However, there are also several disadvantages. First, speech systems often have resource bounds (time and/or space) that make it infeasible to align a large speech file to its transcription. Second, human transcriptions of long speech segments may contain more errors than those produced over shorter segments. Third, forced alignments on longer stretches of speech can produce poorer results (H.Vereecken et al., 1997).

We will investigate the impact of segmentation, acoustic model, and adaptation by force aligning reference transcriptions of a subset of the Blood Pressure Corpus, which is described in detail in section 2.1.. Section 2.2. describes the setup of the alignment experiments and discusses the evaluation metrics used in the evaluations. Finally section 2.3. describes the results of the alignment experiments and draws some conclusions.

2.1. The Blood Pressure Corpus

The Blood Pressure (BP) Corpus consists of conversations between a health care worker and a patient concerning the need for a blood pressure check. In each trial, the health care worker discusses the (hypothetical) elevated blood pressure reading for the patient and tries to persuade him/her to obtain a follow-up reading sometime within the next month. The patient is either easy or hard to convince. The dialogs begin with the subject who plays the health care worker (HCW) informing the subject who plays the patient of his or her (hypothetical) blood pressure measurement is either slightly (in the hard to convince case) or moderately (in the easy to convince case) elevated. The discussions terminate when either the patient agrees to obtain a follow-up reading on a specific day or the health-care worker decides that no further efforts would be persuasive.

The corpus is constructed from interactions involving ten subjects. Two were females in their final year of a four-year undergraduate nursing program who played the role of the health care worker in the scenario. The remaining subjects included four females and four males, ages 18-30, recruited at the University of Wisconsin Milwaukee College of Engineering and Applied Science. These subjects were told that the hypothetical blood pressure reading represented an elevated measurement and that the nurse would try to convince them to return for a follow-up visit and that they should be either relatively easy to convince or

relatively hard to convince (depending on the experimental condition).

In each session, the two subjects sat next to each other and wore blue smocks over their clothing. Six digital video cameras were used. For each participant, one of the cameras was focused on his/her face to capture information about gaze and the other two cameras were calibrated so that once correspondence between points in the two cameras was established, the 3D positions and velocities of his/her hands could be obtained. Using a five foot-wide prism with a known constellation of points, we are able to obtain points with typical average errors within 1 mm in x and y and about 1.5 mm in z (toward the cameras). The maximal errors are within 4 mm. This has been sufficient for measurements involving conversational gesture interaction. We used off-the-shelf consumer-grade miniDV 30 frames-per-second cameras in progressive scan. The audio for each participant was digitally recorded using a Shure Sm94 unidirectional boom mounted microphone that was placed at a distance of eight inches from the subjects' mouths. The video and audio were synchronized using a movie 'clapper' device. The video was digitized on an SGI workstation and saved in SGI MPEG format. The audio was initially sampled at 44.1K and then downsampled to 14,700 KHz for analysis.

Transcriptions for each session were prepared by a professional transcription service. The audio files were then segmented in Praat, and the reference transcript for each segment was stored with the corresponding time interval. This enabled the transcription to be checked (and updated when necessary) by listening to each segment with high quality headphones using the audio channel corresponding to the speaker.

2.2. Data Preparation

To investigate forced alignments in the BP corpus, we selected five speech files from this data set corresponding to five distinct subjects. There was a total of 1035 words to be aligned in these files. Since the speech files in the multimodal corpus are natural conversations recorded in a normal laboratory environment, the background noise is not negligible. Additionally, due to recording conditions there is considerable channel cross talk in the speech files. Table 1 provides pertinent details concerning the speech files we used, in particular, the file name, the corresponding subject identifier, the subject's gender, the experimental condition (easy or hard to convince), the length of the speech file, and the number of words spoken by the subject.

File	ID	Gender	Cond.	Length (sec.)	Words
green-01	HC1	F	H	89.53	214
green-05	HC2	F	E	98.83	239
red-01	P1	F	H	269.74	249
red-02	P2	M	H	188.99	228
red-05	P5	F	H	164.86	105
Total Words					1035

Table 1: This table summarizes attributes of the speech files used for the forced alignment experiments. P stands for patient, HC for health care worker, F for female, M for male, E for the easy-to-convince condition, and H for the hard-to-convince condition.

All forced alignments are scored relative to gold standard reference alignments (RAs). A reference alignment for each speech file was created by hand fixing an initial automatic forced alignment generated by the HTK-WSJ system. The boundaries were adjusted by a Dr. Harper in Praat based on information provided by the speech waveform, the

spectrogram, F0, and energy and by listening to the speech using high quality headphones.

To evaluate the impact of segmentation on alignment accuracy, we segmented for each speech file based on knowledge provided by the manually aligned transcripts. Given the alignment, silences of a half second of silence or more are identified in the speech file. These silences break the file into stretches of speech separated by stretches of silence, as shown in Figure 1. These stretches of speech are then padded with 0.2 seconds of silence at the beginning and end to create the segments that are used for evaluation. Such a segment is represented by the box in Figure 1. Note that padding was consistent for all segments except for those that may not have sufficient silence due to their position at the beginning or end of the speech file. Over the five speech files, the average length of a segment was 3.47 seconds with a standard deviation of 3.04 seconds. The maximum length segment was 14.89 seconds and the minimum length segment was 0.55 seconds.

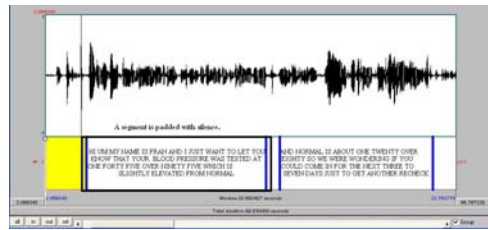


Figure 1: The boxed region depicts a speech segment made up of a stretch of speech padded with 0.2 seconds of silence at its beginning and end. The corresponding reference transcript is associated with that region.

When an alignment is manually repaired, the process can be quite time consuming, especially when word boundaries are shifted far from where they belong. Hence, to evaluate the quality of the automatically generated forced alignments, we chose a metric based on the extent of shift for each word boundary. Since word endings and word beginnings may exhibit different patterns, we chose to measure the distance between the beginning of each word in a forced alignment and the beginning of that word in the reference alignment, as well as the distance at the end of the word. We then calculate the average and standard deviation over the word beginning shifts and the word ending shifts.

2.3. Results

We first compare the alignment accuracy for each ASR system on the speech files in table 1 using either the entire speech file and its transcript or the segmented files and their transcripts. As can be seen in Table 2, forced alignments to segments of speech results in smaller average word boundary shifts over all of the models than forced alignments to entire conversation sides. The boundary shifts for the Aligner tool are much greater than for any of the other systems. Clearly there is an interaction between model type and training material and whether the speech file is segmented or not. Although stronger models are typically better at aligning an entire conversation side than the weaker models, there is a strong advantage to working with segments. It is interesting to note that the speech files used to train the Aligner and HTK systems, i.e., TIMIT and WSJ, respectively, were in fact segmented into spoken sentences, in contrast to the speech files in SWB and BN corpora, which were segmented more arbitrarily.

Figure 2 shows the relative strength of each of the ASR systems when performing the segmented alignments. The curves for the SWB-ISIP, BN-SRI, and SWB-SRI systems are at top and very close to each other. The curve for

the *Aligner* tool shows the worst overall alignment performance; however, since it is trained on such a small training set of read speech and its acoustic model is the most elementary among the systems, this result is not unexpected. It is clear that factors associated with the acoustic models, such as the number of Gaussian mixtures, monophone versus triphone modeling, and training corpus size, seem to have an impact on alignment accuracy, as can be seen by the improvements going from *Aligner* to WSJ-HTK, and from WSJ-HTK to BN-SRI. In all three systems, the preponderance of the training data involves speech produced in a reading task; hence, it is clear that the sophistication of the model and amount of training data is central to the improvement. The SWB-SRI and SWB-ISIP models achieved the best overall performance, with the former achieving a lower standard deviation, possibly due to its more sophisticated acoustic model. BN-SRI performed somewhat worse than the best models, but since it was trained on read speech, this degradation makes sense. The speech in our alignment data set involves spontaneous conversation. To obtain the most accurate word alignments, it makes sense that a high quality acoustic model well trained with conversational speech of a broad range of speakers would perform more strongly.

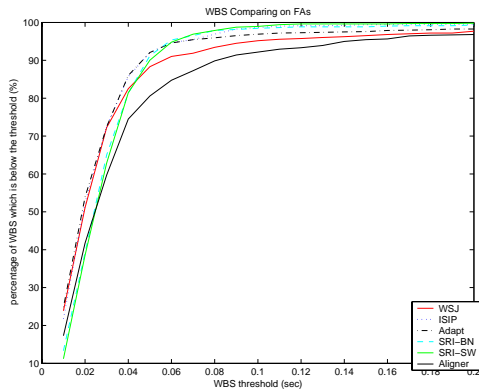


Figure 2: WBS cumulative distribution function plots

Next, we examine the impact of adaptation on the HTK and SRI Decipher models. The average and standard deviation of the word beginning boundary shift and the word ending boundary shift for each adapted system are given in the Table 3. Although speaker adaptation improves the word end boundaries for the HTK model, it does not completely make up for the fact that the model is less tailored to the type of speech being aligned. WSJ-HTK improved more on word endings with adaptation than the SRI mod-

Model	Segment File?	Mean WBBS (SD)	Mean WEBS (SD)
Aligner	no	23,506.1 (26,156.5)	23,432.6 (26,156.5)
Aligner	yes	41.5 (76.0)	44.4 (73.5)
WSJ-HTK	no	2,331.5 (10,273.5)	2,228.7 (9,861.5)
WSJ-HTK	yes	27.9 (41.2)	35.1 (54.2)
BN-SRI	no	35,609.1 (39,606.9)	35,564.0 (39,558.3)
BN-SRI	yes	25.078 (39.579)	30.690 (42.164)
SWB-ISIP	yes	25.9 (52.5)	29.7 (51.7)
SWB-SRI	no	3,901.0 (14,970.9)	4,054.4 (14,925.1)
SWB-SRI	yes	24.3 (27.5)	29.8 (32.99)

Table 2: Average Word Beginning Boundary Shift (WBBS) and Word Ending Boundary Shift (WEBS) in msec and their Standard Deviations (SD) over entire conversation sides or segments.

els, possibly due to the fact that the model was trained on much less data.

Model	Mean WBBS (SD)	Mean WEBS (SD)
WSJ-HTK	29.244 (45.560)	30.233 (48.855)
BN-SRI	24.287 (39.936)	30.603 (44.603)
SWB-SRI	24.112 (27.555)	29.946 (31.681)

Table 3: Average Word Beginning Boundary Shift (WBBS) and Word Ending Boundary Shift (WEBS) in msec and their Standard Deviations (SD) in adapted models given segmented speech files.

A systematic study was conducted to compare several ASR systems on a word forced alignment task. From this study, we found that segmenting the speech files prior to alignment improves the overall alignment accuracy and that alignment accuracy is enhanced by using more advanced acoustic models and more training data matched on speaking style (conversational versus planned) to the data to be aligned. Speaker adaptation improves the models somewhat, but more so for the weaker models.

3. Acknowledgements

This work was supported in part by NSF under award number 9980054, ARDA under contract number MDA904-03-C-1788, DARPA under contract MDA972-02-C-0038, and Purdue Research Foundation. Part of this work was carried out while the last author was on leave at NSF. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of NSF, ARDA, or DARPA. We would like to thank Collin Wightman, Joe Picone, Naveen Parihar, and Andreas Stolcke for their technical support of these experiments.

4. References

Anastasakos, Tasos, John McDonough, Richard Schwartz, and John Makhoul, 1996. A compact model for speaker-adaptive training. In *Proceedings of the Inter. Conf. on Spoken Language Processing*.

Chen, Lei, Mary Harper, and Francis Quek, 2002. Gesture patterns during speech repairs. In *Proc. of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)*. Pittsburgh, PA.

Eickeler, Stefan, Martha Larson, Wolff Rüter, and Joachim Köhler, 2002. Creation of an Annotated German Broadcast Speech Database for Spoken Document Retrieval. In *3rd LREC*. Canary Islands.

Hosom, John Paul, 2000. *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. Ph.D. thesis, Oregon Graduate Institute of Science and Technology.

H. Vereecken, A. Vorstermans, J.P. Martens, and B. Van Coile, 1997. Improving the phonetic annotation by means of prosodic phrasing. In *Proceeding of Euro Speech 1997*.

Kessens, Judith M. and Helmer Strik, 2003. On automatic phonetic transcription quality: lower word error rates do not guarantee better transcriptions. *Computer Speech and Language*.

H. Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond, 2003. The VIC corpus of conversational speech. *IEEE Trans. Acoust., Speech, Signal Processing*.

Quek, Francis, David Mcneill, Robert Bryll, Susan Duncan, Xinfen Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari, 2002. Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction*, 9(3).

Sundaram, A. Ganapathiraju, J. Hamaker, and J. Picone, 2000. ISIP 2000 conversational speech evaluation system. In *Speech Transcription Workshop 2001*. College Park, Maryland.

Stolcke, A., H. Bratt, J. Butzberger, H. Franco, V.R. Rao Gaddi, M. Plache, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, 2000. The SRI March 2000 Hub-5 conversational speech transcription system. In *Proceedings of the NIST Speech Transcription Workshop 2000*. College Park, MD.

Wightman, Colin and David Talkin, 1997. *The Aligner*. Entropic.