Audio-Visual Anticipatory Coarticulation Modeling by Human and Machine

Louis H. Terry[1]*, Karen Livescu*[3]*, Janet B. Pierrehumbert*[2] *and Aggelos K. Katsaggelos*[1]

[1]Northwestern University, Department of EECS, Evanston, IL, 60208, USA
[2]Northwestern University, Department of Linguistics, Evanston, IL, 60208, USA
[3]Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

Abstract

The phenomenon of anticipatory coarticulation provides a basis for the observed asynchrony between the acoustic and visual onsets of phones in certain linguistic contexts. This type of asynchrony is typically not explicitly modeled in audio-visual speech models. In this work, we study within-word audio-visual asynchrony using manual labels of words in which theory suggests that audio-visual asynchrony should occur, and show that these hand labels confirm the theory. We then introduce a new statistical model of audio-visual speech, the asynchrony-dependent transition (ADT) model. This model allows asynchrony between audio and video states within word boundaries, where the audio and video state transitions depend not only on the state of that modality, but also on the instantaneous asynchrony. The ADT model outperforms a baseline synchronous model in mimicking the hand labels in a forced alignment task, and its behavior as parameters are changed conforms to our expectations about anticipatory coarticulation. The same model could be used for speech recognition, although here we consider it only for the task of forced alignment for linguistic analysis.

Index Terms: audio-visual speech recognition, audio-visual asynchrony, anticipatory coarticulation, dynamic Bayesian networks

1. Introduction

Audio-visual anticipatory asynchrony is a naturally occurring linguistic phenomenon in which the visible gestures (mainly the lip gesture) for a speech segment occur in advance of other articulatory components of the segment, so that the visible gestures – the *viseme* – are seen before the corresponding phone is heard. A common example of this is the "pre-rounding" seen in the word "school". The lips begin to round for the /uw/ sound while the /k/ (or even /s/) is still being produced. This phenomenon in known as "anticipatory coarticulation".

Preservatory coarticulation is a similar effect, but instead of one gesture beginning in advance, a gesture continues after. Though anticipatory coarticulation is more pervasive in English, the extent and directionality of coarticulation patterns differ across languages [1, 2].

Anticipatory coarticulation has been studied since at least the 1930s [3]. In 1966, Henke proposed a computer model of the articulation of English stop + vowel with a novel "look-ahead" mechanism for anticipatory coarticulation [4].

In the speech recognition literature, Bregler and Konig showed in [5] that, on average, acoustic features were maximally correlated with visual features 120 ms in the past. This was also reported in psychological experiments by Benoit [6].

In the area of audio-visual speech biometrics, [7] cites these asynchrony effects as one of the major open problems.

command	color*	preposition	letter*	digit*	adverb
bin	blue	at	A-Z	1-9, zero	again
lay	green	by	excluding W		now
place	red	in			please
set	white	with			soon

Table 1: Vocabulary of GRID Corpus

Currently, a typical approach to modeling asynchrony in audio-visual speech is the coupled HMM (CHMM) [8], in which state transitions in each modality depend on the state of the other modality. In this approach, asynchrony is typically allowed only within the boundaries of each phone/viseme, whereas observed asynchrony often crosses multiple phone boundaries. In contrast, the asynchronous dynamic Bayesian network model of [9] allows asynchrony across multiple phones/visemes within a word, but does not account for the asymmetry that is typical to audio-visual asynchrony. Here we develop a model of asynchrony that both spans multiple phones/visemes and allows for explicit modeling of anticipatory coarticulation.

To investigate anticipatory coarticulation, we collected manual labels of phone and viseme onsets in words that are likely to exhibit anticipatory coarticulation. To our knowledge, this kind of study has not been done before. We also develop an audio-visual speech model that can account for both anticipatory and preservatory coarticulation. This model should be able to handle asynchrony in a way that is more psycholinguistically accurate than previous work.

2. Corpus and Utterance Selection

This work uses the freely available GRID Corpus [10], which contains 34 subjects each speaking 1000 utterances in a studio environment. Table 1 enumerates the corpus vocabulary, which contains many opportunities for anticipatory coarticulation both within and across words. Eleven types of within-word phenomena were selected for analysis in seventy utterances over ten speakers, yielding 166 instances of within-word coarticulation. The selected instances, which all involve lip rounding or protrusion gestures, were:

- /uw/ in "blue", "two", "soon", "q", and "u"
- /r/ in "zero", "three", "four", and "r"
- /w/ in "now"
- /ch/ in "h"

3. Human Labeling of Anticipatory Coarticulation

Four undergraduate linguistics students who had completed introductory linguistics classes were recruited to hand label the

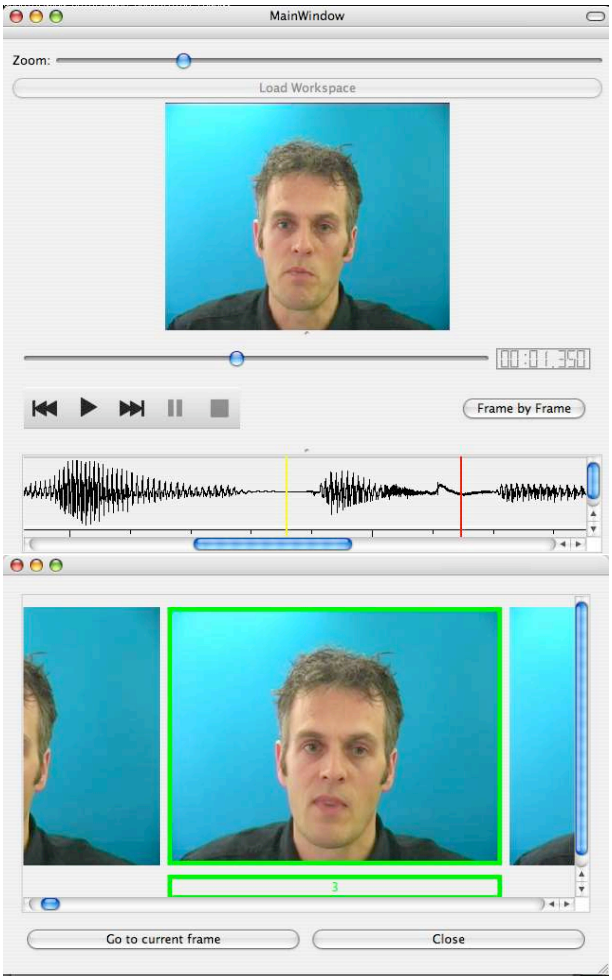

Figure 1: AVDDisplay program created to facilitate easier hand labeling of audio and video onsets.

audio and video onsets of the selected phones/visemes. Labelers were instructed only to "Please label the beginning of the X gesture" (where X is one of the phones above) with no additional instruction, in order to prevent biased labeling. No definition of "beginning" was provided. While this undoubtedly added extra variability to the hand labels, it is a less biased result from which we can draw stronger generalizations. To aid in the hand labeling task, we developed an audio-visual data display tool (AVDDisplay). A screenshot is shown in Figure 1

3.1. Definition of Asynchrony

The video and audio are sampled at different rates and, thus, a convention must be established to define "asynchrony". The video of the GRID Corpus is sampled at a rate of 25 fps, or one frame every 40ms, while the audio is sampled at a much higher rate. In this work, we consider an audio sample to "belong" to a video sample if the audio sample time is within $\pm$20ms of the video sample. In this scheme, each video sample represents the audio samples that precede and succeed it by 20ms.

3.2. Results and Analysis

Upon initial analysis of the results, we noticed that the inter-labeler range of markings for certain words in certain utterances could be fairly large, on the order of 150ms or more. We interpret this as indicative of a particularly difficult word to label and use confidence filtering to exclude such instances. To do so, the ranges of the audio and video labels of all instances of a word were each taken as a sample set and a 99% confidence interval around the mean was calculated. Any word instance for which either modality's range was above the upper bound of that modality's confidence interval was deemed to be of "low confidence" and excluded from further analysis. Even with this filtering, the audio and video ranges averaged approximately 45ms and 75ms with standard deviations of approximately 40ms and 49ms, respectively.

	Median	L1	L2	L3	L4
Avg. Audio Diff.		16.59	-11.16	8.96	-2.50
Avg. Abs. Aud. Diff.		24.63	16.40	17.02	12.70
Avg. Video Diff.		19.78	10.02	-7.81	-15.44
Avg. Abs. Vid. Diff.		31.63	31.55	25.33	22.77
# Early Vid.	35	30	67	13	14
# Early Aud.	13	10	24	36	28
# Synched	72	80	29	71	78

Table 2: Summary statistics for the four labelers L1–L4 and their median **excluding** labels with low confidence. Positive/negative values signify that the mark is before/after the median.

Final markings were derived from the hand labeled data by taking the median of each marking. These median labels yield 35 instances of early video onsets (defined as the video marking occuring more than 20ms before the audio marking), 13 instances of early audio onsets (the audio marking occurred at least 20ms before the video marking), and 72 instances of synchronous onsets (the audio marking was within 20ms of the video marking). Table 2 reports these values as well as the asynchrony breakdown for each labeler and some statistics pertaining to each labeler's performance relative to the median.

The average difference between each labeler's markings and the median reflects the data shown in the labeler's asynchrony breakdown. For instance, labeler 2 marked early video onsets much more often than the median, and this is reflected by the average audio difference being negative (later than the median) and the average video difference being positive (earlier than the median). Averages of absolute differences are also provided in Table 2.

The median markings of the labelers, shown in Table 2, confirm the expectation that video onsets should precede audio onsets more often than the reverse in the chosen words. While the large number of synchronous onsets was not expected, this could be due to the coarseness of the video sampling, which means that only asynchrony greater than 20ms is recognized.

4. Machine Labeling of Anticipatory Coarticulation

To capture the anticipatory coarticulation phenomenon, an audio-visual speech modeling system must be able to allow asynchrony across phone/viseme boundaries. One possible way is to enforce synchrony at word boundaries while letting the audio and visual stream models evolve without constraint within each word. This does not make linguistic sense and, indeed, performs poorly in experiments. To effectively model anticipatory coarticulation, a model must enforce some kind of synchrony constraints.

Our new model is based on the word-synchronous dynamic Bayesian network used in our previous work [11] with the ad-

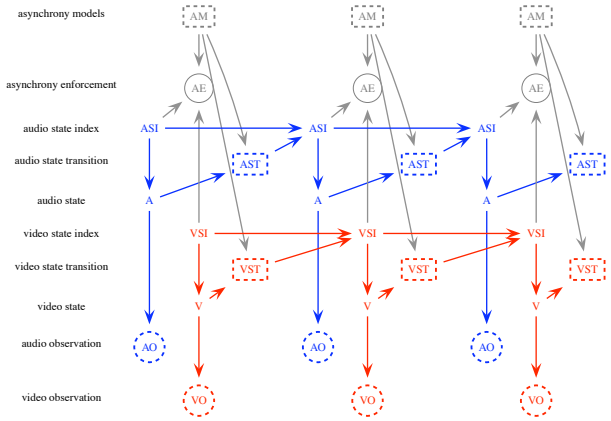

Figure 2: Word-synchronous ADT model for training/alignment. Diagram is simplified for clarity and is conditioned on word-level variables that are not shown.

dition of a synchrony control mechanism based on [9]. This model also takes inspiration from CHMMs [8] in that it allows state transitions to depend on variables other than just the current modality's state. In our case, however, the dependence is on the instantaneous asynchrony rather than the state itself.

4.1. Model Description

Starting with a word-synchronous dynamic Bayesian network model based on our previous work [11], we add an extended version of the asynchrony constraint system of [9]. In [9], the amount of asynchrony is defined as the absolute value of the difference between the state indices of the streams, measured relative to the last synchrony boundary (the beginning of the word). Here we drop the absolute value, which increases the number of parameters in the model but allows us to more correctly model the difference between audio lead and audio lag. This asynchrony model is learned during training. In a CHMM, a modality's state transition probabilities depend on its state as well as the state of the other modality. In our model, the state transition probabilites depend on the modality's state and the amount of instantaneous asynchrony. We hypothesize that when the modalities are asynchronous, they will tend back to synchrony, so the state transition probabilities should be different during asynchrony than during synchrony. We denote the system with asynchrony-dependent transitions as the "ADT" model. Aside from audio-visual stream weights, there are three main parameters of this model: the maximum number of states of audio lag, the maximum nubmer of states of video lag, and the weighting of the asynchrony model.

Figure 2 shows our ADT models as a dynamic Bayesian network. For clarity, state and phone/viseme level variables have been collapsed into single nodes on the graph. Also, some common elements, such as pronunciation variants and stream weighting, are not shown. Blue nodes and edges represent the audio modality, while red nodes and edges represent the video. The grey nodes and edges denote the asynchrony model and its links to the audio and visual modalities. Nodes with no border are deterministic and hidden, while nodes with a circular border are deterministic and observed. Dashed rectangle borders denote hidden, stochastic nodes and dashed circular borders denote observed, stochastic nodes. The observed audio and video input nodes have Gaussian mixture distributions conditioned on their respective state.

A Lag \ V Lag	0		1		2		3	
	Audio	Video	Audio	Video	Audio	Video	Audio	Video
0	43.50	49.00	43.50	44.56	43.50	44.56	43.50	41.22
1	43.50	53.44	43.50	46.78	43.50	44.56	43.50	41.22
2	43.50	52.33	43.50	49.00	43.50	45.67	43.50	43.44
3	42.39	55.67	43.50	49.00	43.50	45.67	43.50	42.33

Table 3: Average absolute audio and video differences between median hand labels and ADT system labels, for instances where hand labels indicate early audio onsets. Smaller numbers imply better performance. Cell (0,0) is the baseline system.

A Lag \ V Lag	0		1		2		3	
	Audio	Video	Audio	Video	Audio	Video	Audio	Video
0	33.97	50.13	32.69	55.27	33.11	61.56	33.77	65.56
1	33.11	43.56	33.97	49.27	33.20	60.41	33.43	54.41
2	33.40	42.99	33.34	48.70	32.86	53.27	32.86	54.41
3	33.11	42.99	33.63	48.70	34.26	54.41	34.00	54.70

Table 4: Average absolute audio and video differences between median hand labels and ADT system labels, for instances where hand labels indicate early video onsets. Smaller numbers imply better performance. Cell (0,0) is the baseline system.

4.2. System Training and A/V Features

All systems were trained using the same technique. First, the audio and visual streams were trained separately. The number of Gaussians was tuned on a development set, using a mixture growing and splitting procedure similar to that of [12]. The single stream models are combined into a multi-stream model and the combined model is refined by iterating twice through the same mixture growing and splitting procedure.

Ten speakers from the corpus were used for these experiments. The training set consisted of 70% of each speaker's utterances and the alignment set contained 10%. The remaining 20% have been reserved for future use. Training used audio and visual stream weights of 0.7 and 0.3, respectively, and were tuned using recognition experiments on the development set.

Audio features are 12 Mel frequency cepstral coefficients plus energy, with delta and acceleration coefficients appended for a total of 39 audio features. Video features are the 90 highest-energy DCT coefficients (corresponding to 95% of the overal energy) of a 60x40 pixel region of interest around the mouth. These coefficients are mapped to a 30-dimensional space using PCA and have their delta and acceleration coefficients appended as well for a final total of 90 visual features.

All models were implemented using the GMTK [13, 14] software package developed at the Univesity of Washington.

4.3. Forced Alignment Experiments

We use the ADT model to perform forced alignment, in which the word sequence is known and the recognizer is responsible for determining the boundaries for all other hidden variables (words, phones, visemes, audio states, video states, etc.). Forced alignment tasks are integral to speech recognition training and database development, and have important roles in scientfic research on speech. Speech recognition systems can be used as forced aligners to generate transcriptions of recordings. This method is less laborious than hand labeling, and it is becoming widespread in experimental research, now that forced alignment has become competitive with hand labeling for some tasks [15].

We compare our forced alignment results to the human labelers by looking at the onset boundaries. For each system,

A Lag \ V Lag	0		1		2		3	
	Audio	Video	Audio	Video	Audio	Video	Audio	Video
0	25.65	25.26	25.55	26.74	25.29	28.06	25.06	30.92
1	25.89	28.58	25.24	29.79	25.72	30.65	25.54	29.81
2	25.48	28.58	25.61	30.98	25.61	29.42	25.59	26.42
3	25.92	30.15	25.32	27.89	25.74	29.39	25.59	28.54

Table 5: Average absolute audio and video differences between median hand labels and ADT system labels, for instances where hand labels indicate synchronous onsets. Smaller numbers imply better performance. Cell (0,0) is the baseline system.

the audio-visual stream weights were determined by optimizing recognition performance over a 1000 utterance development set. The 70 hand labeled utterances come from this development set, so while the stream weights were not chosen completely independent of the labeled utterances, the hand labeled utterances make up a very small portion of the development set.

Tables 3 through 5 show the average absolute differences for the audio and visual streams for the ADT system over various maximum asynchrony constraints and broken down by the median label classification of the utterance. The values in the zero audio/video lag cell of these tables represent a fully synchronous system, a fairly common implementation of an audio-visual speech system and the baseline model against which we can compare. Overall, the audio differences are barely affected by changing the maximum allowed asynchrony, so our analysis will focus on the absolute video differences.

Table 3 shows the model's performance for cases where the median human label indicated an early audio onset. For a fixed audio lag, as allowed video lag increases, performance improves (absolute differences decrease). This is what we would expect as increasing video lag gives the audio more opportunities to precede the video. Conversely, increasing the allowed audio lag for a given video lag decreases performance, as one would expect.

The results for the early video onset cases, shown in Table 4, are also consistent with our linguistic expectations. For a fixed video lag, as allowed audio lag increases, performance tends to improve (absolute differences decrease) until some threshold. Furthermore, the converse holds as well. Again, this agrees with our linguistic intuition about anticipatory coarticulation.

The synchronously labeled cases (Table 5) predictably show the best performance when no asynchrony is allowed, and no significant patterns exist in the rest of the results.

5. Summary and Future Work

In this work, we have studied the labeling of anticipatory coarticulation in audio-visual speech. We have collected a set of manual labels of audio and video phone/viseme onsets, and found that, while the labelers have fairly high variance, their median behavior agrees with our expectations about anticipatory coarticulation. We have also developed a statistical model of audio-visual speech that explicitly accounts for cross-phone, asymmetric asynchrony between the audio and video state streams. Forced alignments with this model show the expected effects of anticipatory coarticulation, given appropriate limits on the allowed lag in each stream. Considering the laborious nature of the manual labeling task, we are optimistic that automatic forced alignment with this type of model can help psycholinguists study audio-visual speech phenomena.

The forced alignment results presented here, while encouraging, depend on setting the appropriate maximum audio/video lag for a given context. This suggests that an asynchrony model that adapts to linguistic context may be needed to more accurately model the effects of anticipatory coarticulation.

In our ongoing work, we are continuing to study forced and manual alignments in the presence of different types of audio-visual asynchrony effects, both for its own sake and for the purpose of improving asynchrony models for audio-visual speech recognition.

6. References

[1] P. S. Beddor and R. A. Krakow, "Perception of coarticulatory nasalization by speakers of English and Thai: evidence for partial compensation," *JASA*, vol. 106, pp. 2868–2887, 1999.

[2] P. S. Beddor, J. Harnsberger, and S. Lindemann, "Acoustic and perceptual characteristics of vowel-to-vowel coarticulation in Shona and English," *Journal of Phonetics*, vol. 30, pp. 591–627, 2002.

[3] P. A. Keating, "Underspecification in phonetics," *Phonology*, vol. 5, no. 2, pp. 275–292, 1988.

[4] W. Henke, "Dynamic articulatory model of speech production using computer simulation," Ph.D. dissertation, MIT, 1966.

[5] C. Bregler and Y. Konig, "Eiegenlips for robust speech recognition," in *Proc. ICASSP*, Adelaide, Austrailia, 1994, pp. II–669–II–672.

[6] C. Benoit, "The intrinsic bimodailty of speech communication and the synthesis of talking faces," *Journal of the Hungarian Telecom. Assoc.*, 1992.

[7] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *IEEE Proceedings*, vol. 94, pp. 2025–2044, November 2006.

[8] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. ICASSP*, vol. 2, 2002, pp. 2013–2016.

[9] K. Saenko and K. Livescu, "An asynchronous DBN for audio-visual speech recognition," *IEEE Spoken Language Technology Workshop*, pp. 154–157, Dec. 2006.

[10] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *JASA*, vol. 120, pp. 2421–2424, 2006.

[11] L. H. Terry and A. K. Katsaggelos, "A phone-viseme dynamic Bayesian network for audio-visual automatic speech recognition," in *Proc. ICPR*, Tampa, FL, Dec. 2008.

[12] J. Bilmes, et al, "Discriminatively structured graphical models for speech recognition," Johns Hopkins University Center for Spoken Language Processing, Tech. Rep., 2001.

[13] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," in *Proc. ICASSP*, Orlando, FL, 2002.

[14] J. Bilmes. The graphical models toolkit. [Online]. Available: http://ssli.ee.washington.edu/ bilmes/gmtk

[15] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proc. of Acoustics*, 2008.