ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

An Agent-based Framework for Auditory-Visual Speech Investigation

Michael Walsh and Stephen Wilson
{michael.j.walsh, stephen.m.wilson}@ucd.ie

Department of Computer Science, University College Dublin, Ireland

ABSTRACT

This paper presents a framework for investigating the relationship between both the auditory and visual modalities in speech. This framework employs intentional agents to analyse multilinear bimodal representations of speech utterances in line with an extended computational phonological model.

1. INTRODUCTION

The broad field of speech technology has seen significant advances in recent years. Nevertheless, state-of-the-art automatic speech recognition (ASR) systems still perform poorly in unrestricted domains (e.g. conversational speech). Typical problems encountered by ASR systems include underspecified input due to noise, and the inability to accurately model coarticulation. Recently some researchers have moved from the traditional phoneme (triphone) based models towards nonlinear phonological models where speech is viewed as multiple tiers of phonological or articulatory features [1][2]. These features can be modelled and extracted autonomously and temporal desynchronisation between features facilitates better modelling of coarticulation. Moreover, given that a single phoneme maps to a number of phonological/articulatory features, significantly less training data is required than in conventional triphone systems. Interestingly, this trend has also recently manifested itself in the Audio-Visual Speech Recognition (AVSR) domain where Saenko et al. [3] demonstrate how visual speech can be viewed in terms of multiple streams of visual features rather than a single string of visemes. The research presented in this paper endorses the view that the use of phonological / articulatory features and their visual counterparts in AVSR systems is worth pursuing, and offers an investigative framework which employs autonomous deliberative reasoning agents to perform audio-visual syllable analysis in line with an extension to a computational phonological model [4][5].

The next section discusses the development of a visual speech gesture set, motivated by the sparse data problem. This is followed by section 3 which introduces the Multi-Agent Time Map Engine (MATE), the agent-based framework at the core of this paper. Section 4 illustrates the operational characteristics of MATE as it parses a bimodal speech representation. The investigative benefits of the framework are then presented in section 5 before concluding with opportunities for future research.

2. SPEECH GESTURES

The gains made by auditory-only ASR systems can, at least in part, be attributed to the widespread availability of high-quality speech corpora that provided repositories of linguistic data for experimentation, training, testing and evaluation. In contrast the number and calibre of similar resources for use in the area of audio-visual speech processing is limited. Consequently, a primary aim of the research presented in this paper is the development of an audio-visual speech corpus, richly annotated in terms of multiple tiers of visual speech gestures, phonemes and syllable boundaries. The CUAVE database [6] provided the underlying video data. The speech contained within it covers a limited domain, being restricted to the first ten ordinal digits and encompassing approximately eighteen phonemes of English. Each utterance was transcribed phonemically by hand with syllable boundaries being inserted.

The work presented here is motivated by a desire to define a visual feature set that can be used to describe the set of visual gestures used in the production of phonemic segments. As a preliminary course of action, five tiers were defined: *dental*, *rounding, spreading, lips,* and *tongue.* Each tier has a number of mutually exclusive visual gesture features, summarised in Table 1.

Tier	Features
Dental	*visible (vis+),* *not-visible (vis-)*
Rounding	*rounding-onset* *round* *rounding-offset* *non-round*
Spreading	*spreading-onset* *spread* *spreading-offset,* *non-spread*
Lips	*open, closed, tucked*
Tongue	*pre-dental* *post-dental* *not-visible (tvis-)*

Table 1

The features for each tier were defined so that they formed a continuous linear gestural description for that tier. Thus the *dental* tier has two features: *visible*, meaning that the teeth are partially or wholly visible, and *non-visible*, indicating that they are completely hidden from view. Similarly, the *lips* tier has features *open* vs *closed.* In addition it has a third feature *tucked* which is used to describe when a speaker's lower lip has been tucked behind the upper teeth, a gesture typically used in the production of labio-dental segments. Two additional tiers *rounding* and *spreading* are used to document the gestures involved in rounding and spreading of the lips during articulation. To this end, each has a set of four features. The sequence of features *rounding-onset to round to rounding-offset* to *non-round* describes the movement of the lips to a rounded and then to a non-rounded state. Likewise, the spreading tier has features which label the movement of the lips to a spread state and then back to being non-spread. Finally, the *tongue* tier has three features which are intended to describe the visibility and relative position of the tongue during speech. The feature *not-visible* is naturally used when the tongue remains hidden from view, while the remaining features *pre-dental* and *post-dental* imply tongue visibility and explicitly reference its position within the mouth. Thus a *pre-dental* tongue would be visible but remain entirely within the mouth. A *post-dental* tongue would again be visible but would protrude past the teeth.

A subset of the audio-visual data contained in the CUAVE database was labelled with respect to this feature set. The result is a richly annotated bimodal speech corpus including full phonemic and syllabic transcriptions as well as a multitiered annotation structure detailing associated visual speech gestures. The continuous nature of the visual annotations (every point on every tier contains a feature), means that it is possible to model segments or phonemes in terms of overlapping combinations of visual speech gestures. This is analogous to considering phonemic segments to be realisations of overlapping phonological phenomena, i.e. a /t/ being a *voiceless*, *dental*, *plosive*. As the tongue, teeth and lips are the most externally visible speech articulators they were identified as being likely candidates for providing important visual information regarding speech and so the five tiers outlined above seek to describe gestural activity for each of them. In order to analyse the visual representations of segments further, mappings between the phonemes and sequences of corresponding visual gestures were automatically identified within the context of learned syllable phonotactics (the permissable combinations of sounds within the syllable domain for a given language, in this case English). The syllable annotations act as the training data to a learning algorithm [7], and a finite state transducer modelling the syllable structure of the initial training corpus is the resulting output. The transducer is an extension of a *phonotactic automaton,* that is a finite state automaton which represents the phonotactics. It contains tapes for each tier of information used within the multilevel annotations, i.e. *dental*, *spreading*, *rounding* etc. Each transition is labelled with a single phonemic segment which can then be mapped onto an equivalent visual representation, either using information from all five visual tiers, or a combination of any number of them. These visual tiers, coupled with five phonological tiers, form the input to the MATE system discussed next.

3. THE MULTI-AGENT TIME MAP ENGINE (MATE)

Given the richly annotated bimodal speech corpus discussed above, a flexible framework is required that facilitates investigation into the merits of the gestures employed as well as the integration of both modalities. This section presents such a framework, known as the *Multi-Agent Time Map Engine* (MATE). The architecture of the framework is illustrated in figure 1. The key components of Time Map recognition are a phonotactic automaton, and a multitiered representation of a speech utterance in terms of phonological features. This representation is then parsed with respect to the phonotactic,

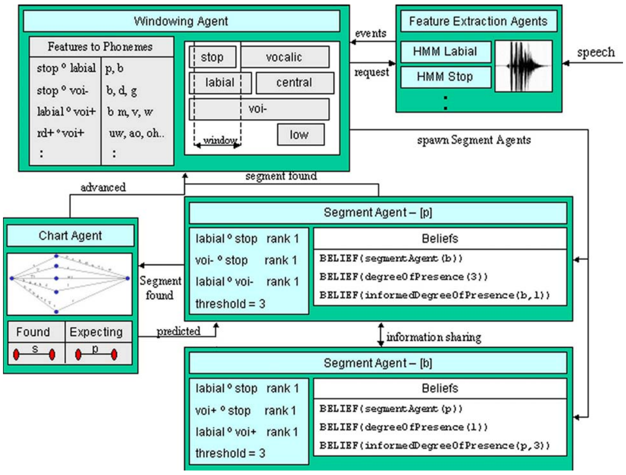

Figure 1. The MATE architecture

automaton. In brief, the phonotactic automaton is used by the parsing algorithm as an anticipatory guide. Essentially, the parser gradually windows through the utterance, guided by the phonotactics searching the multitiered representation to see if the appropriate features for an anticipated segment are present in the current window. The Time Map approach to interpreting multilinear representations of speech utterances has recently been extended [2] in order to investigate the extent to which temporally overlapping phonological features contribute to the identification of a particular segment. This is achieved by attributing rank values to feature overlap relations and a threshold value to segments. In a given window, as feature overlap relations are identified, their rank values are added and if the threshold value for a particular segment is reached the segment is then considered to be present in the window.

Originally developed as a system for investigating multitiered phonological representations of speech utterances in line with the extended Time Map approach discussed above, MATE's flexibility permits its extension to analysis of both phonological and visual speech gestures. This flexibility is achieved through the use of deliberative reasoning agents whose behaviour is governed by high-level behavioural rules.

In brief, agents are software entities that are autonomous, flexible, and are capable of interacting with each other through the use of an Agent Communication Language. In addition they perceive their environment, and are able to affect change upon it. Furthermore, they operate in a goal-directed, self-starting manner. The agents employed here are delivered via a rapid agent prototyping environment known as Agent Factory [8][9]. These agents are equipped with rich mental models. The agent mental state consists of an aggregation of mentalistic attributes which include a belief set, a set of commitments held at a given instance in time, and a set of commitment rules which regulate the adoption of future commitments and beliefs. It is important to note that the agent mental model is transparent to the user.

One key benefit of employing agents to perform interpretation of multilevel representations is that strategies adopted can be altered with ease as the behaviour of agents can be modified at the knowledge level, thus obviating the need to make low-level implementational modifications. MATE employs a number of different agents to parse the bimodal representations mentioned above with respect to a learned finite-state representation of syllable phonotactics. The agent roles employed to deliver syllable analysis are the Chart Agent, Windowing Agent and Segment Agent roles.

3.1 The Chart Agent

This agent performs a number of activities in the syllable identification process. One of the functions of the Chart Agent is to determine all phonotactically anticipated phoneme segments for the current window. The Chart Agent controls the behaviour of at least one finite state phonotactic automaton. In certain cases, however, the Chart Agent can create a copy of the automaton to investigate syllable onset (pre-vowel consonants) and syllable coda (post-vowel consonants) possibilities at the same time. The automata are stochastic, endowed with segment probabilities (based on corpus frequency). The special cases where copies of the automaton are required, and the stochastic nature of the phonotactics, are explained in more detail below. Another activity carried out by the Chart Agent is to monitor progress through the phonotactic automaton.

3.2 The Windowing Agent

The Windowing Agent takes the multilinear representation of a given utterance, and, based on the phonological (or visual) event with the smallest temporal endpoint, creates a window on the utterance. The Windowing Agent then identifies potential segments that may be present based on partial examination of the bimodal feature content of the window. Potential segments can be identified by using a resource which maps features to their respective segments. On the basis of a partial analysis of the feature content of the window the Windowing Agent can spawn multiple Segment Agents, for which there is feature evidence, to perform detailed investigations of the window and attempt to establish their presence in the window.

3.3 The Segment Agent

Each spawned Segment Agent has a number of constraints which it seeks to satisfy by identifying the temporal overlap of relevant features in the window. These constraints are ranked and as each constraint is satisfied its rank value is added to a running total known as a *degree of presence*. If the degree of presence reaches a specific threshold the Segment Agent can consider itself identified. For example, a Segment Agent seeking to identify a /b/ based on phonological information might be aiming to satisfy the following feature overlap constraints:

1. stop ° voiced rank 0.35
2. stop ° labial rank 0.25
3. voiced ° labial rank 0.40
4. threshold = 1.0

where ° represents temporal overlap. Thus, if the Segment Agent satisfied constraints 1 and 2 it would have a degree of presence of 0.6. As temporal overlap is also a reflexive relation it is possible to specify these constraints reflexively e.g.

1. stop ° stop rank 0.2

In this case the rank value indicates the contributory role of a single feature, not the contribution of the overlap of two features as in the examples above. This allows for flexibility in investigation as the impact of single features and overlapping features can both be examined.

For the purposes of illustation the next section indicates how these different agent roles combine to analyse a multitiered representation of a speech utterance.

4. BIMODAL PARSING WITH MATE

Figure 2 presents a bimodal multitiered representation of the syllable (and word) *three* /th r iy/ with respect to five phonological tiers, five visual tiers, and absolute time, coupled with a small fraction (due to space restrictions) of the phonotactic automaton discussed in section 2. It is important to note the coarticulatory nature of the input (features spread and are not necessarily coterminous). With respect to the phonotactics each segment has an associated probability. This probability is acquired during the learning process and reflects the fact that certain segment clusters are more likely than others. In other words certain paths through the automaton have a greater probability of being taken than others. Once MATE is activated the **Platform Viewer** appears which shows all agents active on the platform at any given time. Figure 3 presents three screenshots of the Platform Viewer illustrating the change in the number of agents residing on the platform as MATE examines the first window, **w1,** shown in figure 2. Screenshot (a) indicates the contents of the platform at the beginning of the parse. The first Windowing Agent **WA(1)** and the Chart Agent **CA** are both present and active. The **CA** agent governs the phonotactic automaton and from the initial node, node 0, predicts /th/ and /w/ segments. Screenshot (b) indicates that two new agents have been spawned on the platform. These are the Segment Agents **th(1)** and **dh(1)**, the (1) acting as a window index. These agents are spawned as a result of **WA(1)** 's analysis of **w1**, identifying the temporal overlap of the *den* and *frc* (*fricative*) features as

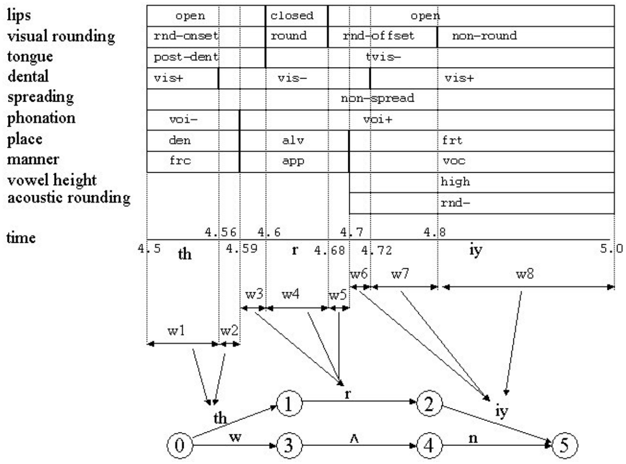

Figure 2. Parsing with MATE

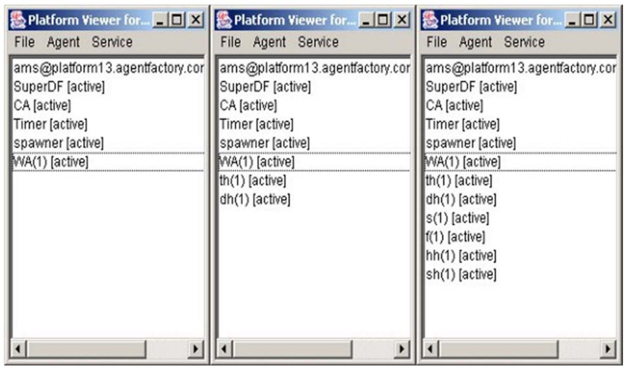

Figure 3. Screenshots of the Platform Viewer, left to right (a), (b), (c).

evidence for a /th/ or /dh/. These Segment Agents immediately seek to determine which of their constituent overlap constraints are present in the window. There was no evidence for **CA**'s prediction of /w/. On its next iteration **WA(1)** identifies the temporal overlap of the *voi-* and *frc* features as evidence for a /s/, /f/, /hh/ and /sh/ segments. As a result **WA(1)** spawns new Segment Agents for each in attempt to determine their presence. These new agents, **s(1)**, **f(1)**, **hh(1)**, and **sh(1)** can be seen in screenshot (c) of figure 3. Each of the Segment Agents examines the feature overlap relations present in the window to determine their respective degrees of presence. In this case the contents of **w1** satisfy all the temporal overlap constraints required by **th(1)**, i.e. this Segment Agent's degree of presence equals its required threshold value. As a result **th(1)** will inform both **WA(1)** and **CA** that it has recognised its segment. The Windowing Agent then instructs the other Segment Agents to terminate, logs the recognised segment and spawns a new windowing agent **WA(2)**. The Chart Agent **CA** accepts /th/ as input and advances the automaton to state 1. Further developments on the platform are not illustrated; however, the parsing continues by **WA(1)** informing **WA(2)** of the resources it will require and adopting a belief to self-terminate:

BELIEF(informed(WA(2),?resources)) **=>**

COMMIT(**Self,Now,Belief**(true),adoptBelief

(**BELIEF**(selfTerminate)))

where **?resources** is a variable storing resource addresses.

The new windowing agent **WA(2)** continues by examining **w2** of the utterance. In this case the coarticulatory nature of the input plays a role. As can be seen in figure 2 the *vis+* feature on the *dental* tier has a shorter duration than the other features associated with the /th/ segment. As a result the *vis-* feature is present in **w2** and Segment Agent analysis of this window yields a /th/ segment which is *below-threshold* as the window is effectively underspecified for /th/. The Chart Agent, **CA**, however is aware that the previous window also yielded a /th/. Consequently **w2** is considered a transition window between steady state segments and **CA** does not advance the automaton. A smoothing operation has essentially taken place.

In a similar fashion **w3** yields below-threshold **/r/**, analysis of **w4** produces a fully-specified /r/, and **w5** offers a below-threshold /r/. Again **CA** is aware of the coarticulatory nature of the input and only advances the automaton once, from node 1 to node 2. Similar processing takes place for windows **w6**, **w7**, and **w8** with respect to /iy/.

As active Segment Agents determine their respective degrees of presence they share this information with other Segment Agents in the community. As a result a Segment Agent is aware of the strength of its competitors and if it believes its degree of presence is too weak it elects to self-terminate. In this way a winning candidate emerges.

In the event that two (or more) competing candidates have equal degrees of presence below their respective thresholds the **CA** can elect to accept both of them, advance the automaton in one direction for the first candidate and make a copy of the automaton and advance in the other direction for the other candidate. At the end of processing the Chart Agent, **CA**, ranks all hypotheses by multiplying out the probabilities of the segments on each path taken.

5. MATE AS AN INVESTIGATIVE FRAMEWORK

In essence MATE employs deliberative reasoning agents to interpret temporally overlapping events using finite-state machinery. The MATE agents perform a number of functions including windowing through the utterance, identifying temporal overlap of both phonological and visual features, decision making with respect to the contributions such overlaps make to phoneme identification, and the consequent identification and ranking of well-formed (and indeed ill-formed) syllables. The resources employed by MATE agents, e.g. mappings from features (both phonological and visual) to phoneme segments, and phonotactic automata specification, are declarative in nature. Consequently MATE is both language and feature set independent, allowing visual speech gestures to be easily incorporated via the extension of resources and the addition of tiers. The investigation of how both modal streams should be integrated is achieved through the use of rankings. This is quite a useful feature as it facilitates prioritisation of particular features (from either modality) over others. For example, in the context of employing MATE as a parser in a bimodal automatic speech recognition system, where visual and phonological events are identified using classifiers (Hidden Markov Models, Artificial Neural Networks etc.) the use of ranks enables MATE to compensate for poor quality feature extraction. That is, if it is known that certain features are particularly difficult to extract from the audio-visual signal then these features can be given

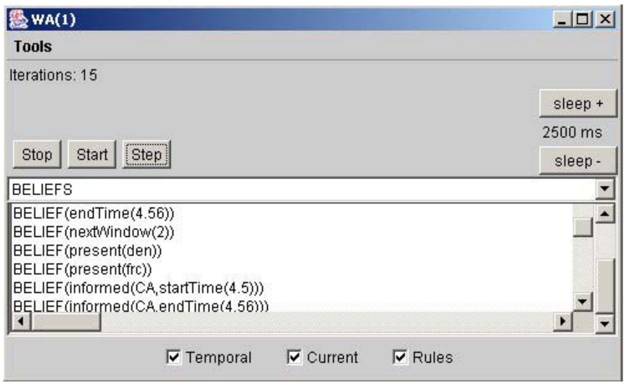

Figure 4. The agent mental state is visible at all times.

low rank values. Similarly features with strong acoustic or visual correlates can be given high rank values, i.e. play a more significant contributory role to the identification of a segment. Hence the use of rankings permits investigation into how best to integrate both modalities.

An additional benefit of using agents is the transparency of their mental models. The beliefs of the MATE agents can be visualised during run-time (see figure 4) and consequently the entire parsing process can be stepped through. In addition parsing strategies can be modified at the knowledge level, i.e. they can be altered using high-level predicate calculus rules. In the parsing example of section 4 MATE parsed the multitiered input without explicit awareness of the distinction between the modalities. It simply sought to identify feature overlaps with respect to a feature-to-segment mapping resource. One alternative strategy would be to equip the agents with beliefs linking features with their respective modality. In this way MATE could for example focus on identifying segments using only phonological features, while employing the visual modality only in cases of significant acoustic underspecification (i.e. Segment Agents with degrees of presence significantly below their thresholds). This kind of change in parsing strategy could be specified entirely at the knowledge level, no low-level implementational changes would have to be made.

6. CONCLUSION

This paper has presented a highly novel framework for investigating the relationship between both the auditory and visual modalities within the domain of the syllable. The MATE framework employs intentional agents, and declarative resources, to analyse multilinear bimodal representations of speech utterances in line with an extended computational phonological model. As a result the framework is, flexible, extensible, and language and feature-set independent.

7. ACKNOWLEDGEMENTS

This material is based on works supported by the Science Foundation Ireland under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

8. REFERENCES

[1]. Sun, J., and Deng, L. “An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition”. *Journal of the Acoustical Society of America*, 111(2):1086-1101, 2002.

[2]. King, S., and Taylor, P. “Detection of Phonological Features in Continuous Speech using Neural Networks”. *Computer Speech and Language*, 14:333-353, 2000.

[3]. Saenko, K., Darrell, T., and Glass, J. "Articulatory Features for Robust Visual Speech Recognition." In: *Proceedings of ICMI*, State College, PA, 152-158;

[4]. Carson-Berndsen, J., and Walsh, M. “Interpreting Multilinear Representations in Speech.” In: *Proceedings of the 8th Australian Conference on Speech Science and Technology*, Canberra, December 2000; 472-477..

[5]. Walsh, M., Kelly, R., Carson-Berndsen, J., O'Hare, G.M.P., and Abu-Amer, T. “A Multi-Agent Computational Linguistic Approach to Speech Recognition.” In: *Proceedings of 18th International Joint Conference on Artificial Intelligence* (IJCAI-03), Acapulco, Mexico. AAAI Press.

[6]. Patterson, E.K., Gurbuz, S., Tufekci, Z., and Gowdy, J.N. "CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research," In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, May 2002.

[7]. Kelly, R., “ A Language Independent Approach to Acquiring Phonotactic Resources for Speech Recognition”, In: *Proceedings of Computational Linguistics in the UK* (CLUK04), Birmingham, UK, 2004.

[8]. Collier, R., “ Agent Factory: A Framework for the Engineering of Agent-Oriented Applications”. PhD thesis, University College Dublin, 2002.

[9]. Collier, R., O’Hare, G.M.P., Lowen, T., and Rooney, C., “ Beyond Prototyping in the Factory of Agents”. In: *Proceedings of the 3rd Central and Eastern European Conference on Multi-Agent Systems (CEEMAS’03)*, Prague, Czech Republic, 2003.