Temporal Alignment using the Incremental Unit Framework

Casey Kennington
Boise State University
Boise, Idaho, U.S.A.
caseykennington@boisestate.edu

Ting Han
Bielefeld University
Bielefeld, Germany
ting.han@uni-bielefeld.de

David Schlangen
Bielefeld University
Bielefeld, Germany
david.schlangen@uni-bielefeld.de

ABSTRACT

We propose a method for temporal alignment–a precondition of meaningful fusion–of multimodal systems, using the incremental unit dialogue system framework, which gives the system flexibility in how it handles alignment: either by delaying a modality for a specified amount of time, or by revoking (i.e., backtracking) processed information so multiple information sources can be processed jointly. We evaluate our approach in an offline experiment with multimodal data and find that using the incremental framework is flexible and shows promise as a solution to the problem of temporal alignment in multimodal systems.

CCS CONCEPTS

• **Computing methodologies** → *Discourse, dialogue and pragmatics*;

KEYWORDS

Multimodal, alignment, incremental, fusion

ACM Reference Format:
Casey Kennington, Ting Han, and David Schlangen. 2017. Temporal Alignment using the Incremental Unit Framework. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3136755.3136769

1 INTRODUCTION

Multimodal fusion requires the joint processing of information from various sources which, for technical reasons, may show different temporal characteristics; i.e., a delay between the actual time of an event and when the information about that event is available [18]. This is further complicated by processing delays (which are often variable) of modules that produce some kind of *late fusion* [28]; i.e., a discrete signal (e.g., speech recognizer). We make an important distinction between multimodal temporal alignment; i.e., ensuring that bits of information which originated at the same time from multiple sensors or processing modules are available to be processed jointly, and fusion; i.e., the actual combining of those bits of information. Such a distinction ensures that whatever performs the fusion actually fuses together information that *should* be based on when the bits of information originated temporally.

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136769

For example, consider a human user who is interacting with a robot equipped with a speech recognizer, an object detector, and a deixis detector. If that human user says *move that one over there*, with the words *that* and *there* accompanied by two distinct corresponding deictic gestures, then in order to understand the intent of the user, the system must align what object was pointed at by the user when the word *that* was uttered, and again when the word *there* was uttered in order for fusion to be meaningful. Any temporal misalignment in the three modalities would result in the robot selecting the wrong object or bringing it to the wrong destination. We propose and explore three possible solutions to the temporal alignment problem:

- Each modality has an *activity detector* that informs the system that information from their respective modalities will be forthcoming. The alignment module can then wait for information from all modalities before acting.
- Information from any modality is acted upon immediately, but recalled and reprocessed when other delayed modalities produce information that originated at the same time.
- A combination of the above two approaches.

This paper contributes a novel approach using the *incremental unit* framework (explained in Section 3) as an amenable framework for the above solutions to the temporal alignment problem. The framework lends itself well to alignment because it allows provisions for aligning many sensors with less need for delays (further explained in Section 4). We show through some preliminary experiments that the system performs as expected (explained in section Section 5), however we leave evaluation within a live, multimodal system that interacts with real users for future work. In the following Section, we explain related work.

2 RELATED WORK

Though somewhat indirectly, we build upon previous general work in multimodal interfaces [19] as well as [3] which defined an interval algebra for time-series overlaps (e.g., event X could take place before event Y, or they could overlap completely or partially, etc.). The evaluation explained in Section 3 is a form of late fusion, which is similar to early work in unification-based fusion [13] and fusion at the semantic level [6, 21, 32].[1] [4, 24] were early approaches to temporal alignment in terms of incremental processing making their work directly related to ours, albeit with a different framework. We refer the reader to [20] for a more in-depth review of relevant multimodal fusion literature than we can provide here.

3 THE IU FRAMEWORK

Incremental systems (i.e., spoken dialogue systems–SDS) process input modalities *incrementally*; that is, they process as much as

[1] [11] makes the case that early fusion (i.e., at the feature level) works better than late fusion–we conjecture that the approach presented here could also be used as a precursor to early fusion.

possible as early as possible (e.g., an incremental speech recognizer would process word by word instead of waiting for silence). It has been shown that human users perceive incremental systems as being more natural than traditional, turn-based SDS [1, 5, 16, 25, 27], offer a more human-like experience [10] and are more satisfying to interact with than non-incremental systems [2]. Psycholinguistic research has also shown that humans comprehend utterances as they unfold and do not wait until the end of an utterance to begin the comprehension process [29, 30], which motivates using an incremental framework for alignment.

The *incremental unit* (IU) framework [23] is a conceptual approach to incremental processing which we build on for alignment. Following [15], the IU framework consists of a network of processing *modules*. A typical module takes input data on its *left buffer*, performs some kind of processing on that data, and produces output on its *right buffer*. The data are packaged as the payload of *incremental units* (IUs) which are passed between modules. The IUs themselves are interconnected via so-called *same level links* (SLL) and *grounded-in links* (GRIN), the former allowing the linking of IUs as a growing sequence, the latter allowing that sequence to convey what IUs directly affect it. Important to this framework and what makes it amenable to alignment is that IUs can be *added*, but can be later *revoked* and replaced in light of new information. Figure 1 shows an example of how a speech recognition module takes an audio signal as input and produces word IUs as output. The IU framework can take advantage of up-to-date information, but have the potential to function in such a way that users perceive as more natural by allowing IUs to be added (thereby acted upon without delay) and revoked, if necessary. In this work, we realize the mechanism that handles multimodal temporal alignment as an IU module, explained in the next section.

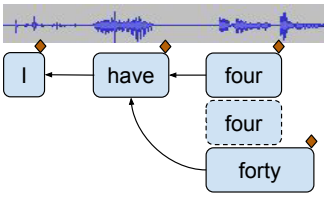

Figure 1: Example of SLL, add, and revoke; the word *four* is added then revoked, being replaced with *forty*. The diamonds denote the point in time when the IU is passed to the next module.

4 IU TEMPORAL ALIGNMENT

For the purposes of this paper, alignment within the IU framework amounts to IUs from multiple modalities arriving at a later processing module (i.e., some kind of fusion module) on its left buffer (i.e., input) simultaneously. In this section, we explain the three proposed solutions for alignment within the IU framework as attempts to meet this goal. For each explanation, we will use examples from two modalities M_1 and M_2 in system aligment IU-module S.

4.1 The Alignment IU-Module

Our model of multimodal temporal alignment is realized as an IU-module which has a left buffer, a right buffer, and a processing element. However, the processing element of this particular module does not perform any additional processing on the IUs that it receives; rather, it serves as a placeholder for those IUs to be passed on, held, or revoked (explained further in the sections that immediately follow). Whereas the left buffer can receive input from multiple modalities, the right buffer should pass those IUs (i.e., produce output on the right buffer) jointly aligned in time. This module assumes that IUs have some kind of timestamp (e.g., IU creation time, or sensor read time) to inform the genesis of an IU to the module. The module can also take in a *threshold* parameter to determine the maximum amount of time gap that is allowed between two IUs of two different modalities to be considered for alignment. When two IUs are considered aligned in time, they are linked together via a GRIN link (see Section 3).

4.2 Activity Detection Driven (AD)

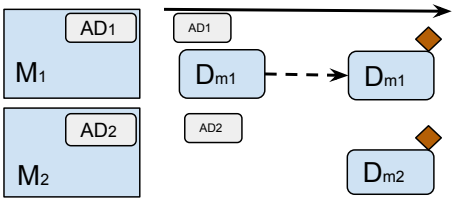

Figure 2: Activity Detection Driven (AD): IUs from each modality's ADs signal that data (i.e., IUs) will be forthcoming. M_1 waits for IUs from other M_2 before being passed along. The dashed arrow denotes the IU was received, but held. The solid arrow indicates time.

For this approach, each modality has a corresponding *activity detector* (e.g., speech recognizers often have voice activity detection) that informs the alignment module that information from their respective modalities will be forthcoming. The alignment module can then wait for information from all modalities which have signaled activity before acting.

Example: S receives an activity detection signal for M_1 and M_2. S then receives information from M_1 at time t_1. M_2 also has data at time t_1, but it has a delay of 250ms. Because of the activity detection signals from both modalities, S waits for information from M_2 (within a specified wait time) before acting. After 250ms, S receives the data from M_2. S outputs the data from M_1 and M_2 simultaneously. This is illustrated in Figure 2. Another example: S receives an activity detection signal from M_1, then receives information from M_1. Because S has no indication that information from M_2 will be forthcoming, it outputs the M_1 information without delay.

Discussion: S is informed by each modality's activity detector. This way S will delay its actions, but only if both modalities will have forthcoming information. S will have less need to revoke, which will produce more informed behaviors. There is also the added difficulty of having some kind of functional activity detector for each modality.[2]

[2]For some modalities, the delay is so small that there is no need for an activity detector.

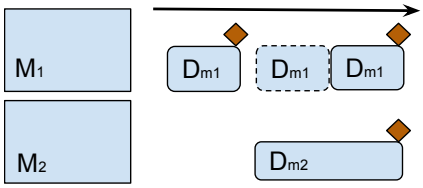

Figure 3: Act and Revoke (AR): IUs from each modality potentially are passed immediately and revoked when information from other modalities is received. Dashed lines denote the point when a sent IU is revoked.

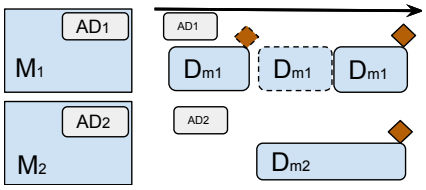

Figure 4: Combined (AR&AD): IUs from each modality can be held for a specified amount of time given a signal from their corresponding ADs, or passed and revoked, if necessary. The dashed diamond denotes a possible output.

4.3 Act and Revoke (AR)

This approach is explained as follows: information from any modality is acted upon immediately, but revoked if other delayed modalities produce information that originated at the same time. This approach makes no use of any activity detection signal.

Example: S receives information from M_1 at time t_1. M_2 also has information at time t_1, but it has a processing delay of 250ms. S proceeds with outputting the information from M_1. However, after 250ms, S receives the information from M_2. S *revokes* information from M_1 and jointly outputs the information from M_1 and M_2. This is illustrated in Figure 3.

Discussion: S should not have to wait for M_2 to act on M_1. This allows S to produce behavior without an unnatural delay. However, in light of new information from M_2, S can revoke and pass information from both modalities jointly.

4.4 Combined AR&AD

This approach combines both AR and AD approaches by treating them as separate IU-modules used serially.

Example: S receives an activity detection signal for M_1 and M_2. S then receives information from M_1 at time t_1. M_2 also has information at time t_1, but it has a processing delay of 250ms. Because of the activity detection signals from both modalities, S waits for information from M_2 for a specified amount of time (say, 150ms) before acting. The information from M_2 has not yet been forthcoming, so S acts upon the information it has from M_1. After 250ms (from t_1), S receives the information from M_2. S *revokes* the processing it has done for M_1 and restarts the processing of the information from M_1 and M_2 jointly. This is illustrated in Figure 4.

Discussion: In this case, the system designer can determine how much delay is acceptable (i.e., AD) so the system produces as informed behavior as possible, but can still get the benefit of AR.

5 EVALUATION

Table 1: Evaluation results for variants AD, AR, and various combinations of AR+AD. The thresholds denote the amount of time (in ms) that was allowed for IUs to be considered aligned; the gaze and speech columns denote the corresponding average delays.

setting	threshold	# revokes	# delays	gaze	speech
none	0	0	0	11.1	309
AD	300	0	593	119.4	397
AR	300	274	0	19.7	402.6
AR+AD	50/300	46	582	119	392
AR+AD	100/300	98	581	118	364
AR+AD	150/300	155	579	120.5	386.3
AR+AD	200/300	196	576	114.9	401.7
AR+AD	250/300	256	490	103.2	410

We implemented each approach as explained in Section 4 as an IU module, denoted *aligner* as part of InproTK [8, 15]. InproTK has been used in several multimodal systems and experiments [12, 14, 17].[3]

Data We apply one multimodal temporally aligned dialogue of the REX corpus [31] using two modalities: speech realized as `AWordIUs` and gaze `GazeIUs` where the gaze payload is an identifier of the object that was being looked at (i.e., processed raw eye tracker data; we used dialogue N2009-N01 form the corpus using the 201 OP-UT IUs as speech and 2286 OP-GZE-N IUs as gaze points). To examine the effectiveness of the alignment as it would perform in a realistic scenario with dynamic and variable delays, we introduce a random delay to the speech where each delay is sampled from a normal distribution (μ=300, σ^2 = 100) in milliseconds.

As an initial sanity check, Figure 5 shows multimodal alignment (in this case, late fusion) between the two modalities using a threshold (i.e., how much time gap is allowed to consider IUs from two modalities for fusion) of 0, 300, and 900 ms. The Alignment Module made no temporal connection between the two modalities when the threshold was set to 0 as expected, some connections when set to 300ms and many connections when set to 900ms. This allows the system designer flexibility in the value of the threshold. This result is similar for all three proposed methods.

Task & Metrics. The goal is minimize delay and maximize stability. Delay is incurred when the aligner holds an IU for some duration, pending arrival of information (i.e., IUs) from another modality. To reflect this, we compute the number of delayed IUs and the average delay in milliseconds for each modality (where delay is the difference between the activity detector IU arrival time and the arrival time of the corresponding IU). To determine stability, we follow [22] and compute a simplified version of *edit overhead*; i.e., how often the model makes unnecessary changes reported as number of revokes. The goal is to minimize all scores. We chose a maximum threshold of 300ms for all tests because it is an expected incremental delay for speech recognition [7] which is also the average artificial delay we add to each speech IU. As above, to add realsitic delay variation, we sample the delay from a normal distribution (μ=300, σ^2 = 100) in milliseconds.

[3] https://bitbucket.org/inpro/inprotk

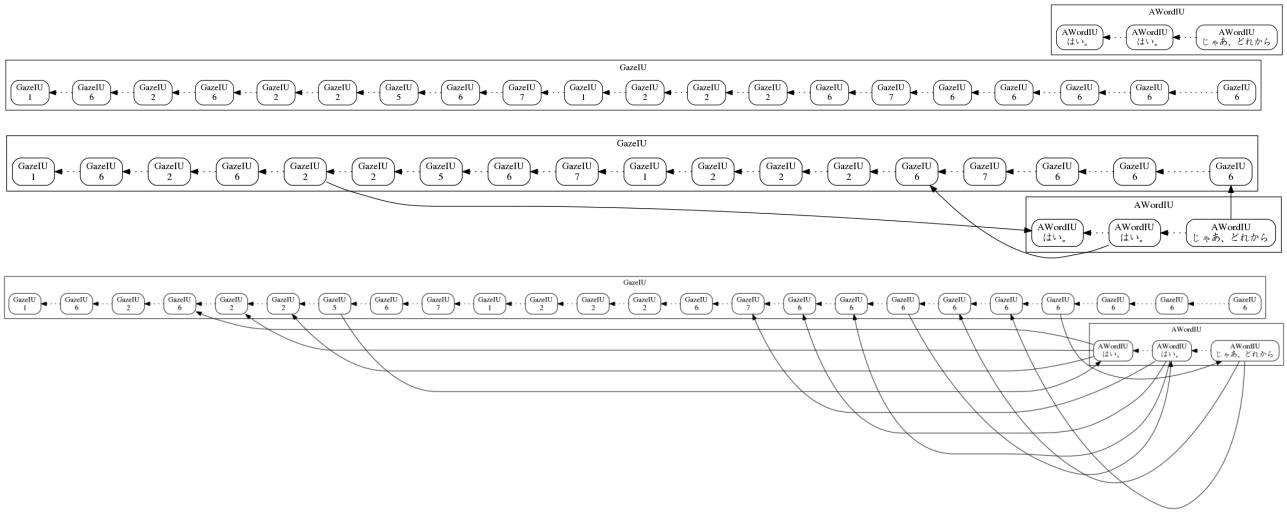

Figure 5: Alignment of two modalities (GazeIU and AWordIU) using the three methods AD, AR, and AR+AD for time thresholds of 0, 300, and 900 respectively from top to bottom. The arrows denote IUs that are considered linked temporally (and hence marked for potential fusion). Higher time thresholds generally means more IUs are linked together temporally.

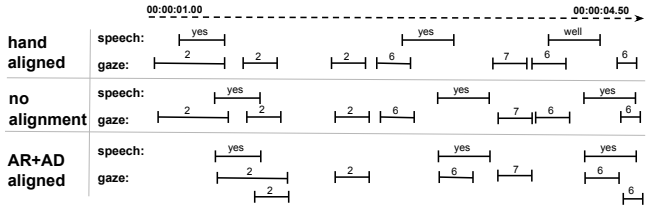

Figure 6: Comparison of hand-aligned vs. non-aligned (where the speech is, on average, 300ms delayed), and AR+AD aligned.

Results. Table 1 shows the results. For the *none* setting, in all cases the speech IUs are 300ms (on average) delayed with no provisions for aligning them. As expected, the average delay for the speech modality was well above the average 300ms delay under all settings. The avg delay gaze difference between *none* and **AR** shows that revoking incurs a very small time overhead resulting in an increased delay. As expected, **AD** produces no `revoke` operations, but causes 593 IUs to be delayed for the full threshold time. **AR** produced 274 revokes, but incurred no additional processing delay. The revokes and processing delays for various settings applying **AR+AD** are shown in the remaining rows. As the threshold for **AR** increased so did the number of revokes. The lowest average delay for speech was **AR+AD** 100/300 with an average delay of 364 ms with 98 revokes and 581 delays. This is an encouraging result; the setting for **AR+AD** 100/300 produced some revokes and delayed some IUs, but not all in all cases, showing that indeed IUs don't always need to be delayed, but if they (wrongly) are sent, they can be revoked.

Figure 6 shows a comparison between the modalities when hand-aligned, when there is no alignment (i.e., an avg 300ms delay for speech) and alignment using the best setting of **AR+AD** (100/300). There is no way to recover the avg 300ms delay in the speech, but the **AR+AD** setting aligns the modalities whereas those alignments would be lost otherwise. This compares to the 300ms middle part in Figure 5. Note that each IU has information about the sensor read timing–the figure portrays the timing of IUs as they begin processing in a system (e.g., a dialogue system).

6 DISCUSSION & CONCLUSION

The results show that the IU framework can be used flexibly for temporal alignment depending on the circumstances; in all cases with the ability to align and fuse multiple modalities. We noted above that alignment as an IU module could potentially be used for realistic, natural behavior. For example, if the **AR** alignment module passes an IU to the next module which, as a result, begins to produce a behavior (e.g., an utterance or a robot begins to reach for and object), only to revoke that IU, the system can produce some kind of disfluency (e.g., by uttering "um" or by stopping the reach) which is seen by human as more natural [9, 26].

In this paper, we have proposed a solution to temporal alignment based on the IU framework. It can flexibly handle any threshold as a parameter and can theoretically handle any number of modalities, though here we only considered two modalities. Moreover, the module can be instantiated anywhere in a system where there may be need for temporal alignment between two modalities or modules. We evaluated our approach systematically using real data and some simple metrics. No approach to alignment can possibly recover the delays caused by the sensors and modules, but as shown here there are several options when handling alignment which could potentially allow a system to produce behavior as soon as possible while benefiting from aligned fusion between modalities. For future work, we will apply this in a live multimodal system that interacts with human users and evaluate how those users perceive the naturalness of the interaction.

Acknowledgements. We appreciate the anonymous reviewers for their comments and suggestions.

REFERENCES

[1] Gregory Aist, James Allen, Ellen Campana, Lucian Galescu, Carlos Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. 2006. Software architectures for incremental understanding of human speech. In *Proceedings of CSLP*. 1922—-1925.

[2] Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, and Mary Swift. 2007. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Pragmatics*, Vol. 1. Trento, Italy, 149–154.

[3] James F. Allen. 2013. Maintaining Knowledge about Temporal Intervals. In *Readings in Qualitative Reasoning About Physical Systems*. 361–372. https://doi.org/10.1016/B978-1-4832-1447-4.50033-X

[4] Afshin Ameri, Batu Akan, Baran Çürüklü, and Lars Asplund. 2011. A General Framework for Incremental Processing of Multimodal Inputs. In *Proceedings of ICMI*. ACM. http://delivery.acm.org/10.1145/2080000/2070521/p225-ameri.pdf?ip=132.178.207.4

[5] Layla El Asri, Romain Laroche, Olivier Pietquin, and Hatim Khouzaimi. 2014. NASTIA: Negotiating Appointment Setting Interface. In *Proceedings of LREC*. 266–271.

[6] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems* 16, 6 (2010), 345–379. https://doi.org/10.1007/s00530-010-0182-0

[7] Timo Baumann. 2013. *Incremental spoken dialogue processing: Architecture and lower-level components*. Ph.D. Dissertation. Bielefeld University.

[8] Timo Baumann and David Schlangen. 2012. The InproTK 2012 release. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*. 29–32.

[9] Simon Betz, Petra Wagner, and David Schlangen. 2015. Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis. *Interspeech 2015* (2015).

[10] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication* 50, 8-9 (2008), 630–645. https://doi.org/10.1016/j.specom.2008.04.002

[11] Hatice Gunes and Massimo Piccardi. 2005. Affect recognition from face and body: early fusion vs. late fusion. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, Vol. 4. IEEE, 3437–3443.

[12] Julian Hough and David Schlangen. 2017. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *Proceedings of the 2017 Conference on Human-Robot Interaction (HRI2017)*.

[13] Michael Johnston, Philip R Cohen, David McGee, Sharon L Oviatt, James A Pittman, and Ira Smith. 1997. Unification-based multimodal integration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 281–288.

[14] Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *Proceedings of SigDial 2013*. 173–182.

[15] Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. InproTKs: A Toolkit for Incremental Situated Processing. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, Philadelphia, PA, U.S.A., 84–88. http://www.aclweb.org/anthology/W14-4312

[16] Casey Kennington and David Schlangen. 2016. Supporting Spoken Assistant Systems with a Graphical User Interface that Signals Incremental Understanding and Prediction State. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Los Angeles, 242–251. http://www.aclweb.org/anthology/W16-3631

[17] Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and Stefan Schlangen. 2014. Situationally Aware In-Car Information Presentation Using Incremental Speech Generation: Safer, and More Effective. In *Proceedings of the Workshop on Dialogue in Motion (DM), EACL 2014*. 68–72. http://pub.uni-bielefeld.de/publication/2663059

[18] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–81.

[19] Sharon Oviatt. 2003. Multimodal interfaces. *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* 14 (2003), 286–304.

[20] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[21] Gerhard Russ, Brian Sallans, and Harald Hareter. 2005. Semantic Based Information Fusion in a Multimodal Interface.. In *CSREA HCI*. Citeseer, 94–102.

[22] David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. In *Proceedings of the 10th SIGdial*. Association for Computational Linguistics, London, UK, 30–37. http://www.ling.uni-potsdam.de/

[23] David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. In *Dialogue & Discourse*, Vol. 2. 83–111. https://doi.org/10.5087/dad.2011.105

[24] Ethan Selfridge and Michael Johnston. 2015. Interact: Tightly-coupling Multimodal Dialog with an Interactive Virtual Assistant. In *Proceedings of ICMI*. ACM. https://doi.org/10.1145/2818346.2823301

[25] Gabriel Skantze and Anna Hjalmarsson. 1991. Towards Incremental Speech Production in Dialogue Systems. In *Word Journal Of The International Linguistic Association*. Tokyo, Japan, 1–8.

[26] Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1–8.

[27] Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09* April (2009), 745–753. https://doi.org/10.3115/1609067.1609150

[28] Cees G M Snoek, Marcel Worring, and Arnold W M Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 399–402.

[29] Michael J. Spivey, Michael K. Tanenhaus, Kathleen M. Eberhard, and Julie C. Sedivy. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology* 45, 4 (2002), 447–481. https://doi.org/10.1016/S0010-0285(02)00503-0

[30] Michael Tanenhaus, Michael Spivey-Knowlton, Kathleen Eberhard, and Julie Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N.Y.)* 268, 5217 (1995), 1632–1634. https://doi.org/10.1126/science.7777863

[31] Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. 2012. The REX corpora : A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*. 422–429.

[32] Minh T Vo and Alex Waibel. 1997. *Modeling and interpreting multimodal inputs: A semantic integration approach*. Technical Report. DTIC Document.