Abstract—We present a number of “meaning” elements carried by possible spoken dialogue texts. All of them are well known within Linguistics, Semantics, Formal Language Disciplines and Philosophy of Language and several formal models have already been given in the past. We also show how any concept of meaning goes far beyond lexicon and “grammar”, involving at least world knowledge, but also attitudes, emotions and intentions.

We finally focus on the role that can be played by visual information in resolving problematic linguistic forms and augmenting knowledge about a human-human interaction scenario. In particular, the acoustical stimuli which persons are exposed to always have correlated visual cues on the face. In addition, we also postulate how visual cues can be useful in disambiguating a dialogue by providing contextual information thus fusing soft and hard data.

Keywords: Soft data, Linguistics, Formal semantics, Pragmatics

I. INTRODUCTION

In this paper we briefly survey the main issues in extracting meaning from speech and the possible solutions that have been proposed in the years within different disciplines. In particular, we point out the richness of the spoken word compared to text as it conveys to the hearer far more information than a mere transcription of the speech.

Let’s focus on a natural language fragment like a dialog or any other linguistic interaction among more than two people. To understand the object of communication and to use the knowledge carried by it, it is necessary to figure out which are the main meaning-bearing, or significant, features or phenomena. Since a long time the significant linguistic and paralinguistic phenomena have been object of investigation based on several theoretical points of view. Each of them focused on a different class of problems.

In this paper we will try to give an account of the main theoretical approaches, showing some of the problems they try to solve. In addition, as text is surpassed by speech in carrying meaning, so does video with respect to the spoken word. In particular, we cast the problem of understanding dialogue in the video surveillance application domain where video information can further help extracting meaning from what is being said by capturing additional information such as face expressions and proximity of the speakers.

Therefore, we hint how the understanding of soft data as expressed in natural language by speakers detected by a video surveillance system can be augmented by combining it with hard data extracted by video sensors. We postulate how the fusion of soft and hard data can help, for intelligence operations, the understanding of the activities and the intentions of detected subjects.

II. GENERAL LINGUISTICS

We all agree now that a language’s important property is organizing its elements into recursive structures, importance fully realized after the 1957 publication of Noam Chomsky’s book Syntactic Structures which presented a formal grammar of a fragment of English. Prior to this, the most detailed descriptions of linguistic systems were of phonological or morphological systems, which tended to be rather closed.

In contemporary linguistics, considering linguistic structures as pairings of meaning and sound, can be recognized different disciplines dealing with different subparts of the linguistic structure, from sound to meaning.

If we leave apart syntax, semantics, pragmatics and discourse analysis, object of formal disciplines which can give us, as we will see, sound means for knowledge extraction, we nevertheless must take into account other areas and concepts that foster significant research, like Phonetics, Phonology and Morphology. Moreover, intersecting with these areas are domains arranged around the different external factors that are considered, we can mention here Stylistics, Language geography, Psycholinguistics and Sociolinguistics.

As anyone would agree that the divisions overlap considerably, we are not trying here to draw borders between disciplines and define them but just mentioning the existence of well founded practices dealing with sound and meaning, some of them giving us information which can be easily (from our perspective and for our purposes, linguists would probably disagree!) integrated into a wider knowledge framework.

We can mention here Phonetics as the study of the physical sounds themselves (phones) of human speech and Phonology which, in contrast to phonetics while grounded on it, is the study of language-specific systems and patterns of sound and gesture.

Phonology deals with sound and gesture units (phonemes) and their possibly different manifestations in phones (allophones - switching allophones of the same phoneme won’t change the meaning of the word, while switching allophones of different phonemes the meaning of the word will change, see Figure 1), the distinctive properties (features - based on articulatory, acoustic, and performance events) which form the basis of meaningful contrast between these units, and their
classification into natural classes based on shared behaviour and phonological processes.

The over a hundred phones recognized by phonetics and the phonological classes are the bases of speech recognition and speaker identification, which is also in the scope of biometrics.

<table>
<thead>
<tr>
<th>utterances</th>
<th>allophones</th>
<th>phonemes</th>
<th>Italian word</th>
</tr>
</thead>
<tbody>
<tr>
<td>['ka:za']</td>
<td>[k]</td>
<td>/k/</td>
<td>casa</td>
</tr>
<tr>
<td>['ha:za']</td>
<td>[h]</td>
<td></td>
<td>in Tuscany</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>utterances</th>
<th>phones</th>
<th>phonemes</th>
<th>English words</th>
</tr>
</thead>
<tbody>
<tr>
<td>['ka:f']</td>
<td>[k]</td>
<td>/k/</td>
<td>calf</td>
</tr>
<tr>
<td>['ha:f']</td>
<td>[h]</td>
<td>/h/</td>
<td>half</td>
</tr>
</tbody>
</table>

Figure 1. In Italian, both phones [k] and [h] belong to the “family” of phoneme /k/, so switching from one to the other the meaning of the uttered word doesn’t change: they are allophones. On the contrary, in English phone [k] belongs to the “family” of phoneme /k/ and [h] belongs to the distinct “family” of phoneme /h/ switching from one to the other the meaning of the uttered word changes: they are not allophones.

Through these linguistic tools, the study of a fragment of conversation can draw a picture of the persons involved in it, making emerge some diversity factors in language:

- **geographical**, with thousands of languages and dialects;
- **cultural**, because the level of education strongly influences the speaking style;
- **physical**, because everyone’s speech organs have slightly different shapes;
- **psychological**, because each person can assume different speaking styles depending on attitudes, emotional states, and intentions.

Although many questions are still unanswered, computational models of speech communication do exist. An important topic - that has recently received considerable attention - is the transmission of emotions in speech communication. Automatic speech recognition (ASR) is an example of a popular field in which the processing of emotions can have a substantial impact and can improve the effectiveness and naturalness of the man-machine interaction. Many of the researches in the field have emphasized the importance of prosodic features (e.g., speech rate, F0 and intensity contours, F0 range) and the importance of the voice quality in the rendering of different emotions in verbal communication [1] [2] [3].

### III. **FORMAL SEMANTICS**

Knowledge of Natural Language involves some capacities. As speakers or writers we must be able to express thought through words, as hearers or readers we must be able to recognize the thoughts expressed from the words we perceive but, above all, to recognize the systematic relations between meaning and linguistic form. Those relations, those rules, somehow guarantee the reconstruction of a thought from a bunch of perceived words.

Formal semantics have been trying since a long time to describe in detail and in a rigorous way the relation meaning-linguistic form. But what is a meaning and what is a linguistic form?

Different uses of the cited terms have developed different theoretical points of view about semantics. *Linguistic form* has ever been the main subject to be focused on since language itself became object of scientific study. It is the topic of what we usually refer to as grammar. Notions of meaning and content are much more problematic.

The perspective according to which the processes of “putting” content “into” linguistic form, or of extracting a content from a linguistic form have something to do with some “language of thought”, has been strongly defended by the psychologist Jerry Fodor [4] with a somehow convincing argument: mental states which have propositional content must be, according to Fodor, *computational*. The process of reasoning cannot be understood unless we assume that beliefs, desires, etc. which act as premises of mental inferences and the conclusions that are drawn from them have some sort of formal, language-like, representational structure within which the particular drawn inference instantiates a general formal inferential pattern defined in terms of the structural relations between premises and conclusion as they appear within that mode of representation.

For thoughts and utterances which concern the actual world there arises the question whether they are true or false. A thought or utterance about the actual world is true if it correctly reflects the way the actual world is, false otherwise. Moreover, truth and falsity are not just any concepts that apply to world-directed utterances and thoughts. Truth play a role of great importance to us, particularly in the context of practical reasoning.

Truth or falsity of a natural language utterance is the product of two independent factors, on the one hand the meaning of the expressions uttered and on the other the factual constitution of its subject matter. A theory explaining the part that meaning plays in determining the truth must succeed in separating these factors. For only then will it enable us to perceive clearly what is being contributed by either.

The method which has thus far proved to be the most effective in achieving this separation is that of model-theoretic semantics. This method was introduced to the study of natural language in the late 60s by Richard Montague [5], who rejected the objection according to which important theoretical differences exist between natural language and formal languages.

---

1. Let’s take the commonsense reading here. It is not within our scope to cover the details of the dispute between the realist and anti-realist conception of truth and epistemic accessibility of facts.
In model-theoretic semantics the subject matter is represented by way of a model, an abstract structure that encodes, in some natural and direct way, the kind of factual information that is pertinent to the truth values of the sentences of the language that is being studied. The object, then, becomes that of articulating, for each sentence \( S \) of this language or language fragment, in which of the possible models \( S \) is true and in which it is false. The interest of such an articulation resides in its details, these details depend on two kinds of structure, on the one hand the structure of the models the theory adopts, and on the other that of the sentences with which it is concerned.

What renders such accounts especially valuable as accounts of meaning is that they make precise how each structural component of a sentence contributes to the determination of the truth values which the sentence assumes in each of the models considered. By specifying what contribution each sentence constituent makes to the truth of the many different sentences in which it occurs as constituent, it tells us also something about the meanings of these constituents. In particular, it will show how the meaning of a complete sentence is connected with the meanings of its constituent parts.

The central ideas which motivate the model-theoretic analysis of linguistic meaning go back to the German mathematician and philosopher Gottlob Frege. It was also Frege’s insight that to explain the meaning of a sentence one must explain under what conditions the sentence is or would be true.

Such a view has been also defended by the philosopher Donald Davidson who, in the late 60s, proposed [6] that a definition of truth for a language structure and the meanings of their component words is about all that a theory of meaning for that language could be asked to deliver.

There exists a theory of meaning which attempts to embed both ideas that the links between language and the world are essential to what constitutes linguistic meaning and that a person has the ability to assign a meaning to the strings of signs or sounds which he reads or hears, and which he recognizes as conforming to the grammar of his language. It is Hans Kamp’s Discourse Representation Theory or DRT [7] [8].

In DRT the interpretations of sentences and fragments of language are constructed in the form of abstract structures, so-called Discourse Representation Structures or DRSs obtained through the application of certain rules to the input sentences, the so-called DRS Construction Rules. A DRS is a couple of sets, on the one hand a set of discourse referents (displayed at the top of the diagram), on the other hand a set of conditions displayed below the universe of discourse, like in the following example \( K_1 \):

\[
\begin{array}{c}
x \\
\text{man}(x) \\
\text{got_in}(x)
\end{array}
\]

In \( K_1 \), “\( x \)” is a discourse referent, “\( \text{man}(x) \)” and “\( \text{got_in}(x) \)” are the DRS-conditions. Ignoring the tense component of the verb, \( K_2 \) is the representation of sentence (1)

1. A man got in.

Discourse referents are useful to represent anaphora. If sentence (1) is followed by another sentence obtaining (2), the interpretation system introduces a new referent which stands for the pronoun “he”, and looks for a suitable referent already present in the DRS to solve it. The resulting structure is the following \( K_2 \):

\[
\begin{array}{c}
x \\
\text{man}(x) \\
\text{got_in}(x) \\
\text{sat_down}(y) \\
\end{array}
\]

2. A man got in. He sat down.

These representations share a comparable structure with the models carried out by model-semantics. Formally they are partial models of little finite domains. The similarity is due to the fact that a sentence ought to incorporate the state-of-the-world conditions to be satisfied for the sentence to be true. Therefore a very natural representation of these conditions can be given by a partial model compatible with a model of the state-of-the-world only when the conditions are satisfied.

Discourse referents are useful to represent anaphora. If sentence (1) is followed by another sentence obtaining (2), the interpretation system introduces a new referent which stands for the pronoun “he”, and looks for a suitable referent already present in the DRS to solve it. The resulting structure is the following \( K_2 \):

\[
\begin{array}{c}
x \\
\text{man}(x) \\
\text{got_in}(x) \\
\text{sat_down}(y) \\
\end{array}
\]

These representations share a comparable structure with the models carried out by model-semantics. Formally they are partial models of little finite domains. The similarity is due to the fact that a sentence ought to incorporate the state-of-the-world conditions to be satisfied for the sentence to be true. Therefore a very natural representation of these conditions can be given by a partial model compatible with a model of the state-of-the-world only when the conditions are satisfied.

Such a theory arose some years ago out of attempts to deal with two distinct puzzling problems of the traditional model-semantics.

The first is the so-called donkey sentences, like “If Pedro owns some donkey, he beats it”, dealing with the (quantifier scope) conflict between the anaphoric connection (indefinite Noun Phrase “some donkey” and pronoun “it” ) and the existential meaning of the word “some”. The second problem deals with tense and aspect and the morphologically different simple past and continuous past in Romance languages.

DRT, through its representation structures, succeeds in giving account of most of the problems (remain open some linguistic forms which can be considered idiosyncratic) and since the publication of [8] more machinery have been added to its tools.

It has been proved to be equivalent to First-Order Calculus [9], it can refer to abstract objects [10], it can be easily interfaced with other structures using ontology systems [11].

IV. SYNTAX, SEMANTICS, PRAGMATICS

The distinctions among syntax, semantics, and pragmatics are due to the American semiotician and philosopher Charles Morris. In [12] Morris distinguished three branches of inquiry within semiotics, the general science of signs: syntax, the study of “the formal relation of signs to one another”, semantics, the study of “the relations of signs to the objects to which the signs are applicable” (their designata), and pragmatics, the study of “the relations of signs to interpreters”. On this view, syntax concerns properties of expressions, such as well-formedness;
semantics concerns relations between expressions and what they are “about”, such as reference and truth-conditions; and pragmatics concerns relations between expressions, their meanings, and their uses in context, such as implicature.

In recent work, many have challenged the autonomy of semantics from pragmatics and the sharp distinction between them implied by the traditional trichotomy. The subdiscipline of formal pragmatics, like in [13], is concerned especially with issues where semantics and pragmatics overlap.

The examples which follow, traditionally included in formal accounts of each of the mentioned branches, actually can be worked out in a more comprehensive formal-semantics framework like DRT.

A. Presupposition

Let’s look at the following example:

(3) Jo’s boss just went to Iraq

It can be argued that (3) carries no meaning at all unless Jo has a boss, but the sentence is not asserting that Jo has a boss, rather it assumes that he exists and the hearer approves the assumption. Information conveyed this way is called a presupposition. A presupposition is some knowledge backgrounded and/or taken for granted, i.e. assumed by the speaker to be already assumed by the hearer to be true. An approximate definition of something called pragmatic presupposition: “A use of a sentence S in a context C pragmatically presupposes a proposition p if p is backgrounded and taken for granted by the speaker in C”.

(4) At 2 o’clock John started to work.

Presupposition: At some time before 2 o’clock, John wasn’t working.

Assertion: At some time after 2 o’clock, John was working.

According to [14] presuppositions in many respects behave as anaphors. A consequence of his presuppositions-as-anaphors view is that the notorious projection problem for presuppositions can be reduced to the problem of resolving anaphoric pronouns.

In the following example the consequent contains a presupposition trigger, the word “boss”, and the triggered presupposition is explicitly stated in the antecedent of the conditional, then the presupposition is blocked.

(5) If Jo has a boss, then Jo’s boss is foreigner.

That sentence doesn’t imply that Jo has a boss, whether in example (6) the presupposition is not stated in the antecedent, so it is allowed to project, i.e. the sentence does imply that Jo has a boss.

(6) If it’s already 4am, then Jo’s boss is probably angry.

Van der Sandt shows how presuppositions can be handled using the same mechanism which resolves anaphoric pronouns in DRT. There is one important difference between pronouns and “real” presuppositions: when no suitable, accessible antecedent can be found for a presupposition, and the presupposition has sufficient descriptive content, it can be accommodated and, so to speak, create its own antecedent.

B. Implicatures

In [15] Grice means “implicate” to cover the family of uses of “imply”, “suggest”, “mean” while things that follow from what a sentence literally “says” or asserts are called “entailments”; so the major distinction Grice draws in his work is between (semantic) entailments, dealing with truth-conditional content and (pragmatic) conversational implicatures.

So, e.g., in (7) what B implied, suggested, or meant is distinct from what B said. All B said was that C had not been to prison yet, but it conversationally implicates that C may have a tendency toward criminal behavior.

(7) A: How is C getting along in his new job at the bank?
   B: Oh, quite well, I think; he likes his colleagues, and he hasn’t been to prison yet.

Before Grice it was widely held that there are considerable mismatches between the standard interpretations of the standard connectives and operators of logic (¬, ¬¬, ∧, ∨, →, “∀x”, “∃x”) and the meanings of their closest counterparts in ordinary Natural Language (“not”, “and”, “or”, “if - then”, “every”, “some” (or “at least one”), ‘the’).

According to Grice the meanings of the connectives and operators of standard logic are much closer to the meanings of their natural language counterparts than had been assumed. He argued that it was a failure to distinguish between semantics and pragmatics, a failure to distinguish between the literal semantic content of a sentence (“what is literally said by a sentence”) and a variety of further kinds of inferences that may reasonably be drawn from the speaker’s use of that sentence in a particular context. A speaker may succeed in communicating (intentionally or unintentionally) much more than what is literally said by the words of her sentence.

(8) a. Jeff earned a lot of money and started his own side business.
   b. Jeff started his own side business and earned a lot of money.

(9) Tests proved that Jones was the author of the document and
   a. he was sent to jail.
   b. he got a promotion.

The sentential conjunction “and” appears to be unambiguous: lexical semantics should specify that its truth-conditional meaning is just the meaning of the logical conjunction “and”. But the sentences convey something else which can be explained within pragmatics, using the concept of conversational implicatures which follow some principles that generate them, Grice’s “Conversational maxims”.

Conversational partners normally recognize a common purpose or common direction in their conversation, and at any
point in a conversation, certain “conversational moves” are judged suitable or unsuitable for accomplishing their common objectives. The general principle, called Cooperative Principle says: Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

Under this very general principle, Grice distinguishes four categories of maxims, characteristic of conversation as a cooperative activity (Maxims of Quantity, Maxims of Quality, Maxim of Relation, Maxims of Manner) with a Quality Supermaxim: Try to make your contribution one that is true.

Some other examples show different implicatures generated by violation of Maxims.

Letter of recommendation: it is “suggested” that the letter writer does not have a very high opinion of Mr. X.

(10) “Dear Sir, Mr. X’s skills are excellent, and his attendance at training sessions has been regular. Yours, etc.”

Metaphor: it is “suggested” that the words are not to be taken in their usual literal sense.

(11) That throws some light on the question.

A “generalized implicature”. Almost any use of a sentence of the form (12) would normally implicate that the person to be met was not Tom’s wife, mother, or sister. (Similarly with other indefinites - “Tom went into a house” implicates that it was not Tom’s house.)

(12) Tom is meeting a woman this evening.

In the following example from [16], despite the violations, hearer is expected to recognize what is happening:

(13) A asks: Where’s Bill?
    B answers: There’s a yellow VW outside Sally’s house.

¿From a third-party point of view, given the Supermaxim, i.e. B telling the truth, a successfully concluded conversation like the previous can be used to derive deductively/abductively different chunks of implicit knowledge:

(14) a. Bill owns a yellow VW
    b. Bill (at least) knows Sally
    c. Both A and B know a. and b.

Recently an account of implicatures within the DRT framework has been given in [17].

C. Stress and intonation

Stress and intonation in languages have been commonly regarded as mere “stylistic factors” which do not contribute to the essential meaning of sentences. In [18] Ray Jackendoff began to construct an account of the semantic effects of some phonological phenomena and tried to show how they fit into a possible theory of discourse. One of them is focus.

Roughly, it is a theoretical notion introduced by linguists [19] to describe and try to explain a systematic correlation between accent and discourse context. Prosodic prominence (here indicated by square brackets labelled with F) is standardly described as marking focus.

The concept of focus is quite obscure but the phenomenon is evident in the following Question/Answer examples (Free Focus) being a. an odd answer:

(15) Who did you introduce Bill to?
    a. # I introduced [Bill]_{F} to Sue.
    b. I introduced Bill to [Sue]_{F}.

Accents also have an impact on truth conditions in sentences with particles like only, even, too, (Association with Focus phenomenon)

(16) a. I only introduced Bill to [Sue]_{F}.
    b. I only introduced [Bill]_{F} to Sue.

Relation between intonation and meaning is also commonly assumed to be mediated by syntax as it is shown in the following examples:

(17) George only broke the VASE.
    a. [George only broke [the vase]_{F}]
       (narrow focus - Noun Phrase)
       meaning “George didn’t break anything else’
    b. [George only [broke the vase]_{F}]
       (broad focus - Verb Phrase)
       meaning “George didn’t do anything else’

Also for the phenomenon of Focus, among the others, has been already given an account [20] within the DRT framework.

D. Indexicals and Demonstratives

We are not going to give any theoretical or philosophical account of these linguistic expressions, which actually are very complex, but it is worth mentioning them. Indexicals are those expressions whose reference shifts from context to context. Some examples are “I”, “here”, “now”, “today”, “he”, “she”, and “that”. Two speakers who utter a single sentence that contains an indexical may say different things. “He”, “his”, “she”, and “her” are sometimes used like bound variables in formal languages. For example, the occurrence of “he” in (18) (on the relevant understanding) functions like an occurrence of a variable that is bound by the occurrence of the quantifier phrase “every man”. Similarly, “her” in (19) (under the appropriate reading) is bound by “every girl”.

(18) Every man believes that he is smart.
(19) Every girl looks after her suitcase.

But there are the uses of these pronouns in which we are interested, the indexical (or demonstrative or deictic) uses, as in (20) and (21).

(20) He likes Ferrari cars [pointing at John], but he does not [pointing at James].
(21) His car is dirty [pointing at John], but his car is clean [pointing at James].
It is quite evident that it is impossible to bind the variable introduced by the pronouns in (20) and (21) unless we see the speaker’s pointing gestures or the speaker’s intention to refer to a particular object.

The same happens with demonstratives related to space and time like “here”, “now”, “the man sitting under the tree”.

V. LINGUISTIC PERFORMANCE, PROXEMICS

Relative spacing among human bodies and posture have been studied [21] as unintentional or habitual reactions to sensory shifts, such as changes in the sound and pitch of a person’s voice. The space around a human being is no neutral space: physical distance among people is correlated with social distance. Physical space can have a protective function and makes us communicate. Each person’s space (sphere) reveals one’s social position, one’s personality, different kinds of interpersonal relationships (intimacy, commitment, etc.). Four main spheres are envisaged as illustrated in Figure 2.

Different cultures maintain different standards of personal space (Figure 3). In Latin cultures, for instance, relative distances are shorter, and people tend to be more comfortable standing close to each other; in Nordic and Japanese cultures the opposite is true. Realizing and recognizing these body spacing and posture can add information about cultural roots of the speakers.

Proxemics also has been defined through different factors in nonverbal communication, or proxemic behaviour categories, that apply to people engaged in conversation. Within intimate and personal sphere, for instance, each person can perceive heat and odour from the other; positions of one person’s shoulders and another’s shoulders (see Figure 4) can vary according to attitude to encourage (4.b.) or discourage (4.a.) communication. These categorizations can very easily transferred in a domain specific ontology but some of the features are evidently difficult to track.

Figure 4. Relative positions of shoulders: the so called sociopetal-sociofugal axis.

VI. “VISUAL” SPEECH

In the context of face-to-face communication, speech is more than the transmission of an acoustical signal. The production of speech sounds is related to very specific and stable geometrical configurations of the lips and the jaw. Human beings are constantly exposed to both the acoustical stimuli and their visual correlates on the face. We comprehend the sense conveyed by the verbal communication by means of the spatio-temporal coherence between the sounds of speech and the facial gestures that are served to partially “shape” those sounds [22], [23].
In particular, the motion of the face conveys insights on speakers' emotions, semantic and phonetic details. In this sense, the measurement of this motion is a prerequisite for any further analysis of its functional characteristics or information content. Several systems exist that track face motion by means of active (OPTOTRAK, Northern Digital Inc.) or passive (ELITE, BTS, Milan, Italy; QUALISYS, Qualisys Medical AB) markers placed directly on the face.

Thus marker-based systems have several advantages (they are spatially very accurate, having sufficient temporal resolution, and returning instantly accessible and processable data). Unfortunately, they have severe limitations, too:

- the necessary equipment is very expensive and highly specialised (it cannot be used in real life, outside the laboratories);
- the systems are invasive (the markers must be attached to the speakers' skin, see Figure 5);
- marker placement requires a-priori decisions about proper measurement locations (it can restrict or bias further analyses).

It is clear that video-based methods will not be able to compete, in the near future, with marker-based methods in terms of resolution and reliability, but it should be evident that video-based face motion analysis bring advantages able to overcome the disadvantages listed in the point (a), (b), and (c) (see above). As an example, at Human Information Science Laboratories (Kyoto, Japan) and at MARCS Auditory Laboratories (University of Western Sydney), Munhall et al. [24] presented a system for video-based analysis of face motion during speech. It consists of an algorithm to measure face motion from standard video recordings by deforming the surface of an ellipsoidal mesh (initialized manually) fit to the face. The method returns measurement points globally distributed over the facial surface.

Within the analysis of video data it is possible to quantify vocal and consonant labial targets of a speech performed with basic emotional patterns (anger, fear, distress, disgust, happiness, and surprise) through the analysis of a number of parameters: Lip Opening, Upper and Lower Lip vertical displacements, Lip Rounding, Anterior/posterior movements (Protrusion) of Upper Lip and Lower Lip (ULP and LLP), Left and Right Corner horizontal displacements (LCX and RCX), Left and Right Corner vertical displacements (LCY and RCY). These parameters can be very useful for word sense disambiguation in automatic speech recognition. In fact, recent studies [25]–[27] show that:

1. the measured values of all the targets of any parameter under investigation in emotional speech utterances are significantly different from the values obtained measuring non-emotional speech utterances
2. some parameters, i.e. Lip Opening, within the range of measured data variation, retain phonetic-phonological distinctions, while others, i.e. labial corner displacements and asymmetries, vary according to different emotional patterns.

VII. DISCUSSION

The arguments accounted for in the previous sections show how difficult and complex can be a formal and robust analysis of linguistic data, particularly as to spoken texts, dialogues and interactions among a number of persons. Several are the problems left open, which we all hope will find a proper solution in the future, but some of them, namely the indexicals, some use of descriptions, the proxemics approach, cannot be dealt with just within a linguistics approach (we mean here any of the general linguistics, semiotics, formal semantics, formal pragmatics approaches).

Most of the utterances which need further knowledge except the linguistic one and refer to a state-of-affairs (spatial and temporal), contextual to the proximities of the speaker and the hearer, can nevertheless be successfully accounted for using contextual video information. It is the case of the following example:

(22) Look at the girl standing beside the vending machine.

It is unfortunately true that it would be almost impossible to get the correct information from (22) uttered in scenarios like the ones depicted in Figure 6 without visual information. In particular, recent advancements in automatic video surveillance systems (see [28] and [29] for recent surveys), allow a sufficient level of situational awareness at least for what concerns the presence of individuals, their movements, and their interactions.

For example, the situation illustrated by Figure 6(a) could be assessed by a multisensor video surveillance system by locating three individuals, one of them positioned near the vending machine (according to a detailed map of the structure). In addition, the availability of multiple sensors would allow face detection algorithms [29] to detect the direction that the two men involved in dialogue (22) are facing.

Even more challenging is the interpretation of metaphors. In English it is absolutely clear the meaning of

(23) It is a bowling ball.

with respect to

(24) He is a bowling ball.
because of the pronoun. But in Italian, as the use of a pronoun is not necessary, the translation (25) could be a good reading for both Figures 7a and 7b.

(25) È una palla da biliardo.

Therefore, it would be difficult to get the correct meaning out of (25) if not properly contextualized. Again, the fusion of additional cues would be needed to disambiguate the sentence. The location of speaker and hearer could be one of such cues.

Finally, proxemics is in the field of paralinguistic codes, so by definition something else than linguistic and textual. Video is the only medium which carries information about (intentional and unintentional) physical behaviour of interacting people.

VIII. CONCLUSIONS AND FUTURE WORK

We tried to show a number of possible “meaning” elements carried by dialogue and which can be used to produce a better understanding of it. All of them are well known within Linguistics and Philosophy of Language and several formal models have been given; but, as far as we know not much effort has been put in trying to fuse them.

In addition, speech understanding can be improved by exploiting additional cues as those provided by video sensors. In particular, several ambiguities, intentionally or unintentionally conveyed by the speaker, can be resolved by analysing the context in which the dialogue is taking place. We therefore envisage the fusion of soft and hard data as a possible way to improve the automatic understanding of oral communications.

Future work includes the following. First, it should be “measured” the actual significance of these features in the overall knowledge extraction from HUMINT sources. Second, the different formal treatment methods should be analysed trying an harmonization. Third, a framework for fusing soft and hard data should be identified.

REFERENCES


