# *Who is "You"?* Combining Linguistic and Gaze Features to Resolve Second-Person References in Dialogue*

***Matthew Frampton***[1]***, Raquel Fernández***[1]***, Patrick Ehlen***[1]***, Mario Christoudias***[2]***,***
***Trevor Darrell***[2] and ***Stanley Peters***[1]

[1]Center for the Study of Language and Information, Stanford University
{frampton, raquelfr, ehlen, peters}@stanford.edu

[2]International Computer Science Institute, University of California at Berkeley
cmch@icsi.berkeley.edu, trevor@eecs.berkeley.edu

## Abstract

We explore the problem of resolving the second person English pronoun *you* in multi-party dialogue, using a combination of linguistic and visual features. First, we distinguish generic and referential uses, then we classify the referential uses as either plural or singular, and finally, for the latter cases, we identify the addressee. In our first set of experiments, the linguistic and visual features are derived from manual transcriptions and annotations, but in the second set, they are generated through entirely automatic means. Results show that a multimodal system is often preferable to a unimodal one.

## 1 Introduction

The English pronoun *you* is the second most frequent word in unrestricted conversation (after *I* and right before *it*).[1] Despite this, with the exception of Gupta et al. (2007b; 2007a), its resolution has received very little attention in the literature. This is perhaps not surprising since the vast amount of work on anaphora and reference resolution has focused on text or discourse - mediums where second-person deixis is perhaps not as prominent as it is in dialogue. For spoken dialogue pronoun resolution modules however, resolving *you* is an essential task that has an important impact on the capabilities of dialogue summarization systems.

Besides being important for computational implementations, resolving *you* is also an interesting and challenging research problem. As for third person pronouns such as *it*, some uses of *you* are not strictly referential. These include discourse marker uses such as *you know* in example (1), and generic uses like (2), where *you* does not refer to the addressee as it does in (3).

(1) It's not just, you know, noises like something hitting.

(2) Often, you need to know specific button sequences to get certain functionalities done.

(3) I think it's good. You've done a good review.

However, unlike *it*, *you* is ambiguous between singular and plural interpretations - an issue that is particularly problematic in multi-party conversations. While *you* clearly has a plural referent in (4), in (3) the number of its referent is ambiguous.[2]

(4) I don't know if you guys have any questions.

When an utterance contains a singular referential *you*, resolving the *you* amounts to identifying the individual to whom the utterance is addressed. This is trivial in two-person dialogue since the current listener is always the addressee, but in conversations with multiple participants, it is a complex problem where different kinds of linguistic and visual information play important roles (Jovanovic, 2007). One of the issues we investigate here is

*We thank the anonymous EACL reviewers, and Surabhi Gupta, John Niekrasz and David Demirdjian for their comments and technical assistance. This work was supported by the CALO project (DARPA grant NBCH-D-03-0010).

[1]See e.g. http://www.kilgarriff.co.uk/BNC_lists/

[2]In contrast, the referential use of the pronoun *it* (as well as that of some demonstratives) is ambiguous between NP interpretations and discourse-deictic ones (Webber, 1991).

*Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 273–281,
Athens, Greece, 30 March – 3 April 2009. 

how this applies to the more concrete problem of resolving the second person pronoun *you*.

We approach this issue as a three-step problem. Using the AMI Meeting Corpus (McCowan et al., 2005) of multi-party dialogues, we first discriminate between referential and generic uses of *you*. Then, within the referential uses, we distinguish between singular and plural, and finally, we resolve the singular referential instances by identifying the intended addressee. We use multimodal features: initially, we extract discourse features from manual transcriptions and use visual information derived from manual annotations, but then we move to a fully automatic approach, using 1-best transcriptions produced by an automatic speech recognizer (ASR) and visual features automatically extracted from raw video.

In the next section of this paper, we give a brief overview of related work. We describe our data in Section 3, and explain how we extract visual and linguistic features in Sections 4 and 5 respectively. Section 6 then presents our experiments with manual transcriptions and annotations, while Section 7, those with automatically extracted information. We end with conclusions in Section 8.

## 2 Related Work

### 2.1 Reference Resolution in Dialogue

Although the vast majority of work on reference resolution has been with monologic text, some recent research has dealt with the more complex scenario of spoken dialogue (Strube and Müller, 2003; Byron, 2004; Arstein and Poesio, 2006; Müller, 2007). There has been work on the identification of non-referential uses of the pronoun *it*: Müller (2006) uses a set of shallow features automatically extracted from manual transcripts of two-party dialogue in order to train a rule-based classifier, and achieves an F-score of 69%.

The only existing work on the resolution of *you* that we are aware of is Gupta et al. (2007b; 2007a). In line with our approach, the authors first disambiguate between generic and referential *you*, and then attempt to resolve the reference of the referential cases. Generic uses of *you* account for 47% of their data set, and for the generic *vs.* referential disambiguation, they achieve an accuracy of 84% on two-party conversations and 75% on multi-party dialogue. For the reference resolution task, they achieve 47%, which is 10 points over a baseline that always classifies the next speaker as the addressee. These results are achieved without visual information, using manual transcripts, and a combination of surface features and manually tagged dialogue acts.

### 2.2 Addressee Detection

Resolving the referential instances of *you* amounts to determining the addressee(s) of the utterance containing the pronoun. Recent years have seen an increasing amount of research on automatic addressee detection. Much of this work focuses on communication between humans and computational agents (such as robots or ubiquitous computing systems) that interact with users who may be engaged in other activities, including interaction with other humans. In these situations, it is important for a system to be able to recognize when it is being addressed by a user. Bakx et al. (2003) and Turnhout et al. (2005) studied this issue in the context of mixed human-human and human-computer interaction using facial orientation and utterance length as clues for addressee detection, while Katzenmaier et al. (2004) investigated whether the degree to which a user utterance fits the language model of a conversational robot can be useful in detecting system-addressed utterances. This research exploits the fact that humans tend to speak differently to systems than to other humans.

Our research is closer to that of Jovanovic et al. (2006a; 2007), who studied addressing in human-human multi-party dialogue. Jovanovic and colleagues focus on addressee identification in face-to-face meetings with four participants. They use a Bayesian Network classifier trained on several multimodal features (including visual features such as gaze direction, discourse features such as the speaker and dialogue act of preceding utterances, and utterance features such as lexical clues and utterance duration). Using a combination of features from various resources was found to improve performance (the best system achieves an accuracy of 77% on a portion of the AMI Meeting Corpus). Although this result is very encouraging, it is achieved with the use of manually produced information - in particular, manual transcriptions, dialogue acts and annotations of visual focus of attention. One of the issues we aim to investigate here is how automatically extracted multimodal information can help in detecting the addressee(s) of *you*-utterances.

| Generic | Referential | Ref_Sing. | Ref_Pl. |
|---|---|---|---|
| 49.14% | 50.86% | 67.92% | 32.08% |

Table 1: Distribution of *you* interpretations

| $L_1$ | $L_2$ | $L_3$ |
|---|---|---|
| 35.17% | 30.34% | 34.49% |

Table 2: Distribution of addressees for singular *you*

## 3 Data

Our experiments are performed using the AMI Meeting Corpus (McCowan et al., 2005), a collection of scenario-driven meetings among four participants, manually transcribed and annotated with several different types of information (including dialogue acts, topics, visual focus of attention, and addressee). We use a sub-corpus of 948 utterances containing *you*, and these were extracted from 10 different meetings. The *you*-utterances are annotated as either *discourse marker*, *generic* or *referential*.

We excluded the *discourse marker* cases, which account for only 8% of the data, and of the *referential* cases, selected those with an AMI addressee annotation.[3] The addressee of a dialogue act can be `unknown`, a single meeting participant, two participants, or the whole audience (three participants in the AMI corpus). Since there are very few instances of two-participant addressee, we distinguish only between singular and plural addressees. The resulting distribution of classes is shown in Table 1.[4]

We approach the reference resolution task as a two-step process, first discriminating between plural and singular references, and then resolving the reference of the singular cases. The latter task requires a classification scheme for distinguishing between the three potential addressees (listeners) for the given *you*-utterance.

In their four-way classification scheme, Gupta et al. (2007a) label potential addressees in terms of the order in which they speak after the *you*-utterance. That is, for a given *you*-utterance, the potential addressee who speaks next is labeled 1, the potential addressee who speaks after that is 2, and the remaining participant is 3. Label 4 is used for group addressing. However, this results in a very skewed class distribution because the next speaker is the intended addressee 41% of the time, and 38% of instances are plural - the remaining two classes therefore make up a small percentage of the data.

We were able to obtain a much less skewed class distribution by identifying the potential addressees in terms of their position in relation to the current speaker. The meeting setting includes a rectangular table with two participants seated at each of its opposite longer sides. Thus, for a given *you*-utterance, we label listeners as either $L_1$, $L_2$ or $L_3$ depending on whether they are sitting opposite, diagonally or laterally from the speaker. Table 2 shows the resulting class distribution for our dataset. Such a labelling scheme is more similar to Jovanovic (2007), where participants are identified by their seating position.

## 4 Visual Information

### 4.1 Features from Manual Annotations

We derived per-utterance visual features from the Focus Of Attention (FOA) annotations provided by the AMI corpus. These annotations track meeting participants' head orientation and eye gaze during a meeting.[5] Our first step was to use the FOA annotations in order to compute what we refer to as Gaze Duration Proportion (GDP) values for each of the utterances of interest - a measure similar to the "Degree of Mean Duration of Gaze" described by (Takemae et al., 2004). Here a GDP value denotes the proportion of time in utterance $u$ for which subject $i$ is looking at target $j$:

$$GDP_u(i,j) = \sum_j T(i,j)/T_u$$

were $T_u$ is the length of utterance $u$ in milliseconds, and $T(i,j)$, the amount of that time that $i$ spends looking at $j$. The gazer $i$ can only refer to one of the four meeting participants, but the target $j$ can also refer to the white-board/projector screen present in the meeting room. For each utterance then, all of the possible values of $i$ and $j$ are used to construct a matrix of GDP values. From this matrix, we then construct "Highest GDP" features for each of the meeting participants: such

[3] Addressee annotations are not provided for some dialogue act types - see (Jovanovic et al., 2006b).

[4] Note that the percentages of the referential singular and referential plural are relative to the total of referential instances.

[5] A description of the FOA labeling scheme is available from the AMI Meeting Corpus website `http://corpus.amiproject.org/documentations/guidelines-1/`

| For each participant P$i$ |
|---|
| – target for whole utterance |
| – target for first third of utterance |
| – target for second third of utterance |
| – target for third third of utterance |
| – target for -/+ 2 secs from *you* start time |
| – ratio 2nd hyp. target / 1st hyp. target |
| – ratio 3rd hyp. target / 1st hyp. target |
| – participant in mutual gaze with speaker |

Table 3: Visual Features

features record the target with the highest GDP value and so indicate whom/what the meeting participant spent most time looking at during the utterance.

We also generated a number of additional features for each individual. These include firstly, three features which record the candidate "gazee" with the highest GDP during each third of the utterance, and which therefore account for gaze transitions. So as to focus more closely on where participants are looking around the time when *you* is uttered, another feature records the candidate with the highest GDP -/+ 2 seconds from the start time of the *you*. Two further features give some indication of the amount of looking around that the speaker does during an utterance - we hypothesized that participants (especially the speaker) might look around more in utterances with plural addressees. The first is the ratio of the second highest GDP to the highest, and the second is the ratio of the third highest to the highest. Finally, there is a highest GDP *mutual gaze* feature for the speaker, indicating with which other individual, the speaker spent most time engaged in a mutual gaze.

Hence this gives a total of 29 features: seven features for each of the four participants, plus one mutual gaze feature. They are summarized in Table 3. These visual features are different to those used by Jovanovic (2007) (see Section 2). Jovanovic's features record the number of times that each participant looks at each other participant during the utterance, and in addition, the gaze direction of the current speaker. Hence, they are not highest GDP values, they do not include a mutual gaze feature and they do not record whether participants look at the white-board/projector screen.

### 4.2 Automatic Features from Raw Video

To perform automatic visual feature extraction, a six degree-of-freedom head tracker was run over each subject's video sequence for the utterances containing *you*. For each utterance, this gave 4 sequences, one per subject, of the subject's 3D head orientation and location at each video frame along with 3D head rotational velocities. From these measurements we computed two types of visual information: participant gaze and mutual gaze.

The 3D head orientation and location of each subject along with camera calibration information was used to compute participant gaze information for each video frame of each sequence in the form of a gaze probability matrix. More precisely, camera calibration is first used to estimate the 3D head orientation and location of all subjects in the same world coordinate system.

The gaze probability matrix is a $4 \times 5$ matrix where entry $i, j$ stores the probability that subject $i$ is looking at subject $j$ for each of the four subjects and the last column corresponds to the white-board/projector screen (i.e., entry $i, j$ where $j = 5$ is the probability that subject $i$ is looking at the screen). Gaze probability $G(i, j)$ is defined as

$$G(i, j) = G_0 e^{-\alpha_{i,j} 2/\gamma 2}$$

where $\alpha_{i,j}$ is the angular difference between the gaze of subject $i$ and the direction defined by the location of subjects $i$ and $j$. $G_0$ is a normalization factor such that $\sum_j G(i, j) = 1$ and $\gamma$ is a user-defined constant (in our experiments, we chose $\gamma = 15$ degrees).

Using the gaze probability matrix, a $4 \times 1$ per-frame mutual gaze vector was computed that for entry $i$ stores the probability that the speaker and subject $i$ are looking at one another.

In order to create features equivalent to those described in Section 4.1, we first collapse the frame-level probability matrix into a matrix of binary values. We convert the probability for each frame into a binary judgement of whether subject $i$ is looking at target $j$:

$$H(i, j) = \beta G(i, j)$$

$\beta$ is a binary value to evaluate $G(i, j) > \theta$, where $\theta$ is a high-pass thresholding value - or "gaze probability threshold" (GPT) - between 0 and 1.

Once we have a frame-level matrix of binary values, for each subject $i$, we compute GDP values for the time periods of interest, and in each case, choose the target with the highest GDP as the candidate. Hence, we compute a candidate target for the utterance overall, for each third of the utterance, and for the period -/+ 2 seconds from the

*you* start time, and in addition, we compute a candidate participant for mutual gaze with the speaker for the utterance overall.

We sought to use the GPT threshold which produces automatic visual features that agree best with the features derived from the FOA annotations. Hence we experimented with different GPT values in increments of 0.1, and compared the resulting features to the manual features using the *kappa* statistic. A threshold of 0.6 gave the best *kappa* scores, which ranged from 20% to 44%.[6]

## 5 Linguistic Information

Our set of discourse features is a simplified version of those employed by Galley et al. (2004) and Gupta et al. (2007a). It contains three main types (summarized in Table 4):

— *Sentential features* (1 to 13) encode structural, durational, lexical and shallow syntactic patterns of the *you*-utterance. Feature 13 is extracted using the AMI "Named Entity" annotations and indicates whether a particular participant is mentioned in the *you*-utterance. Apart from this feature, all other sentential features are automatically extracted, and besides 1, 8, 9, and 10, they are all binary.

— *Backward Looking (BL)/Forward Looking (FL) features* (14 to 22) are mostly extracted from utterance pairs, namely the *you*-utterance and the BL/FL (previous/next) utterance by each listener $L_i$ (potential addressee). We also include a few extra features which are not computed in terms of utterance pairs. These indicate the number of participants that speak during the previous and next 5 utterances, and the BL and FL speaker order. All of these features are computed automatically.

— *Dialogue Act (DA) features* (23 to 24) use the manual AMI dialogue act annotations to represent the conversational function of the *you*-utterance and the BL/FL utterance by each potential addressee. Along with the sentential feature based on the AMI Named Entity annotations, these are the only discourse features which are not computed automatically. [7]

[6]The fact that our gaze estimator is getting any useful agreement with respect to these annotations is encouraging and suggests that an improved tracker and/or one that adapts to the user more effectively could work very well.

[7]Since we use the manual transcripts of the meetings, the transcribed words and the segmentation into utterances or dialogue acts are of course not given automatically. A fully automatic approach would involve using ASR output instead of manual transcriptions— something which we attempt in Section 7.

| | |
|---|---|
| (1) | # of *you* pronouns |
| (2) | you (say\|said\|tell\|told\| mention(ed)\|mean(t)\| sound(ed)) |
| (3) | `auxiliary` you |
| (4) | `wh-word` you |
| (5) | you guys |
| (6) | if you |
| (7) | you know |
| (8) | # of words in *you*-utterance |
| (9) | duration of *you*-utterance |
| (10) | speech rate of *you*-utterance |
| (11) | 1st person |
| (12) | general case |
| (13) | person Named Entity tag |
| (14) | # of utterances between *you*- and BL/FL utt. |
| (15) | # of speakers between *you*- and BL/FL utt. |
| (16) | overlap between *you*- and BL/FL utt. (binary) |
| (17) | duration of overlap between *you*- and BL/FL utt. |
| (18) | time separation between *you*- and BL/FL utt. |
| (19) | ratio of words in *you*- that are in BL/FL utt. |
| (20) | # of participants that speak during prev. 5 utt. |
| (21) | # of participants that speak during next 5 utt. |
| (22) | speaker order BL/FL |
| (23) | dialogue act of the *you*-utterance |
| (24) | dialogue act of the BL/FL utterance |

Table 4: Discourse Features

## 6 First Set of Experiments & Results

In this section we report our experiments and results when using manual transcriptions and annotations. In Section 7 we will present the results obtained using ASR output and automatically extracted visual information. All experiments (here and in the next section) are performed using a Bayesian Network classifier with 10-fold cross-validation.[8] In each task, we give raw overall accuracy results and then F-scores for each of the classes. We computed measures of *information gain* in order to assess the predictive power of the various features, and did some experimentation with Correlation-based Feature Selection (CFS) (Hall, 2000).

### 6.1 Generic *vs.* Referential Uses of *You*

We first address the task of distinguishing between generic and referential uses of *you*.

**Baseline.** A majority class baseline that classifies all instances of *you* as referential yields an accuracy of 50.86% (see Table 1).

**Results.** A summary of the results is given in Table 5. Using discourse features only we achieve an accuracy of 77.77%, while using multimodal

[8]We use the the BayesNet classifier implemented in the Weka toolkit `http://www.cs.waikato.ac.nz/ml/weka/`.

| Features | Acc | F1-Gen | F1-Ref |
|---|---|---|---|
| Baseline | 50.86 | 0 | 67.4 |
| Discourse | 77.77 | 78.8 | 76.6 |
| Visual | 60.32 | 64.2 | 55.5 |
| MM | 79.02 | 80.2 | 77.7 |
| Dis w/o FL | 78.34 | 79.1 | 77.5 |
| MM w/o FL | 78.22 | 79.0 | 77.4 |
| Dis w/o DA | 69.44 | 71.5 | 67.0 |
| MM w/o DA | 72.75 | 74.4 | 70.9 |

Table 5: Generic *vs.* referential uses

(MM) yields 79.02%, but this increase is not statistically significant.

In spite of this, visual features do help to distinguish between generic and referential uses - note that the visual features alone are able to beat the baseline ($p < .005$). The listeners' gaze is more predictive than the speaker's: if listeners look mostly at the white-board/projector screen instead of another participant, then the *you* is more likely to be referential. More will be said on this in Section 6.2.1 in the analysis of the results for the singular *vs.* plural referential task.

We found sentential features of the *you*-utterance to be amongst the best predictors, especially those that refer to surface lexical properties, such as features 1, 11, 12 and 13 in Table 4. Dialogue act features provide useful information as well. As pointed out by Gupta et al. (2007b; 2007a), a *you* pronoun within a question (e.g. an utterance tagged as `elicit-assess` or `elicit-inform`) is more likely to be referential. Eliminating information about dialogue acts (w/o DA) brings down performance ($p < .005$), although accuracy remains well above the baseline ($p < .001$). Note that the small changes in performance when FL information is taken out (w/o FL) are not statistically significant.

### 6.2 Reference Resolution

We now turn to the referential instances of *you*, which can be resolved by determining the addressee(s) of the given utterance.

#### 6.2.1 Singular *vs.* Plural Reference

We start by trying to discriminate singular *vs.* plural interpretations. For this, we use a two-way classification scheme that distinguishes between individual and group addressing. To our knowledge, this is the first attempt at this task using linguistic information.[9]

[9]But see e.g. (Takemae et al., 2004) for an approach that uses manually extracted visual-only clues with similar aims.

**Baseline.** A majority class baseline that considers all instances of *you* as referring to an individual addressee gives 67.92% accuracy (see Table 1).

**Results.** A summary of the results is shown in Table 6. There is no statistically significant difference between the baseline and the results obtained when visual features are used alone (67.92% *vs.* 66.28%). However, we found that visual information did contribute to identifying some instances of plural addressing, as shown by the F-score for that class. Furthermore, the visual features helped to improve results when combined with discourse information: using multimodal (MM) features produces higher results than the discourse-only feature set ($p < .005$), and increases from 74.24% to 77.05% with CFS.

As in the generic *vs.* referential task, the white-board/projector screen value for the listeners' gaze features seems to have discriminative power - when listeners' gaze features take this value, it is often indicative of a plural rather than a singular *you*. It seems then, that in our data-set, the speaker often uses the white-board/projector screen when addressing the group, and hence draws the listeners' gaze in this direction. We should also note that the ratio features which we thought might be useful here (see Section 4.1) did not prove so.

Amongst the most useful discourse features are those that encode similarity relations between the *you*-utterance and an utterance by a potential addressee. Utterances by individual addressees tend to be more lexically cohesive with the *you*-utterance and so if features such as feature 19 in Table 4 indicate a low level of lexical similarity, then this increases the likelihood of plural addressing. Sentential features that refer to surface lexical patterns (features 6, 7, 11 and 12) also contribute to improved results, as does feature 21 (number of speakers during the next five utterances) - fewer speaker changes correlates with plural addressing.

Information about dialogue acts also plays a role in distinguishing between singular and plural interpretations. Questions tend to be addressed to individual participants, while statements show a stronger correlation with plural addressees. When no DA features are used (w/o DA), the drop in performance for the multimodal classifier to 71.19% is statistically significant ($p < .05$). As for the generic *vs.* referential task, FL information does not have a significant effect on performance.

| Features | Acc | F1-Sing. | F1-Pl. |
|---|---|---|---|
| Baseline | 67.92 | 80.9 | 0 |
| Discourse | 71.19 | 78.9 | 54.6 |
| Visual | 66.28 | 74.8 | 48.9 |
| MM* | 77.05 | 83.3 | 63.2 |
| Dis w/o FL | 72.13 | 80.1 | 53.7 |
| MM w/o FL | 72.60 | 79.7 | 58.1 |
| Dis w/o DA | 68.38 | 78.5 | 40.5 |
| MM w/o DA | 71.19 | 78.8 | 55.3 |

Table 6: Singular *vs.* plural reference; * = with Correlation-based Feature Selection (CFS).

| Features | Acc | F1-$L_1$ | F1-$L_2$ | F1-$L_3$ |
|---|---|---|---|---|
| MC baseline | 35.17 | 52.0 | 0 | 0 |
| Discourse | 60.69 | 59.1 | 60.0 | 62.7 |
| Visual | 65.52 | 69.1 | 63.5 | 64.0 |
| MM* | 80.34 | 80.0 | 82.4 | 79.0 |
| Dis w/o FL | 52.41 | 50.7 | 51.8 | 54.5 |
| MM w/o FL | 66.55 | 68.7 | 62.7 | 67.6 |
| Dis w/o DA | 61.03 | 58.5 | 59.9 | 64.2 |
| MM w/o DA | 73.10 | 72.4 | 69.5 | 72.0 |

Table 7: Addressee detection for singular references; * = with Correlation-based Feature Selection (CFS).

#### 6.2.2 Detection of Individual Addressees

We now turn to resolving the singular referential uses of *you*. Here we must detect the individual addressee of the utterance that contains the pronoun.

**Baselines.** Given the distribution shown in Table 2, a majority class baseline yields an accuracy of 35.17%. An off-line system that has access to future context could implement a next-speaker baseline that always considers the next speaker to be the intended addressee, so yielding a high raw accuracy of 71.03%. A previous-speaker baseline that does not require access to future context achieves 35% raw accuracy.

**Results.** Table 7 shows a summary of the results, and these all outperform the majority class (MC) and previous-speaker baselines. When all discourse features are available, adding visual information does improve performance (74.48% *vs.* 60.69%, $p < .005$), and with CFS, this increases further to 80.34% ($p < .005$). Using discourse or visual features alone gives scores that are below the next-speaker baseline (60.69% and 65.52% *vs.* 71.03%). Taking all forward-looking (FL) information away reduces performance ($p < .05$), but the small increase in accuracy caused by taking away dialogue act information is not statistically significant.

When we investigated individual feature contribution, we found that the most predictive features were the FL and backward-looking (BL) speaker order, and the speaker's visual features (including mutual gaze). Whomever the speaker spent most time looking at or engaged in a mutual gaze with was more likely to be the addressee. All of the visual features had some degree of predictive power apart from the ratio features. Of the other BL/FL discourse features, features 14, 18 and 19 (see Table 4) were more predictive. These indicate that utterances spoken by the intended addressee are often adjacent to the *you*-utterance and lexically similar.

## 7 A Fully Automatic Approach

In this section we describe experiments which use features derived from ASR transcriptions and automatically-extracted visual information. We used SRI's Decipher (Stolcke et al., 2008)[10] in order to generate ASR transcriptions, and applied the head-tracker described in Section 4.2 to the relevant portions of video in order to extract the visual information. Recall that the Named Entity features (feature 13) and the DA features used in our previous experiments had been manually annotated, and hence are not used here. We again divide the problem into the same three separate tasks: we first discriminate between generic and referential uses of *you*, then singular *vs.* plural referential uses, and finally we resolve the addressee for singular uses. As before, all experiments are performed using a Bayesian Network classifier and 10-fold cross validation.

### 7.1 Results

For each of the three tasks, Figure 7 compares the accuracy results obtained using the fully-automatic approach with those reported in Section 6. The figure shows results for the majority class baselines (MCBs), and with discourse-only (Dis), and multimodal (MM) feature sets. Note that the data set for the automatic approach is smaller, and that the majority class baselines have changed slightly. This is because of differences in the utterance segmentation, and also because not all of the video sections around the *you* utterances were processed by the head-tracker.

In all three tasks we are able to significantly outperform the majority class baseline, but the visual features only produce a significant improve-

[10] Stolcke et al. (2008) report a word error rate of 26.9% on AMI meetings.

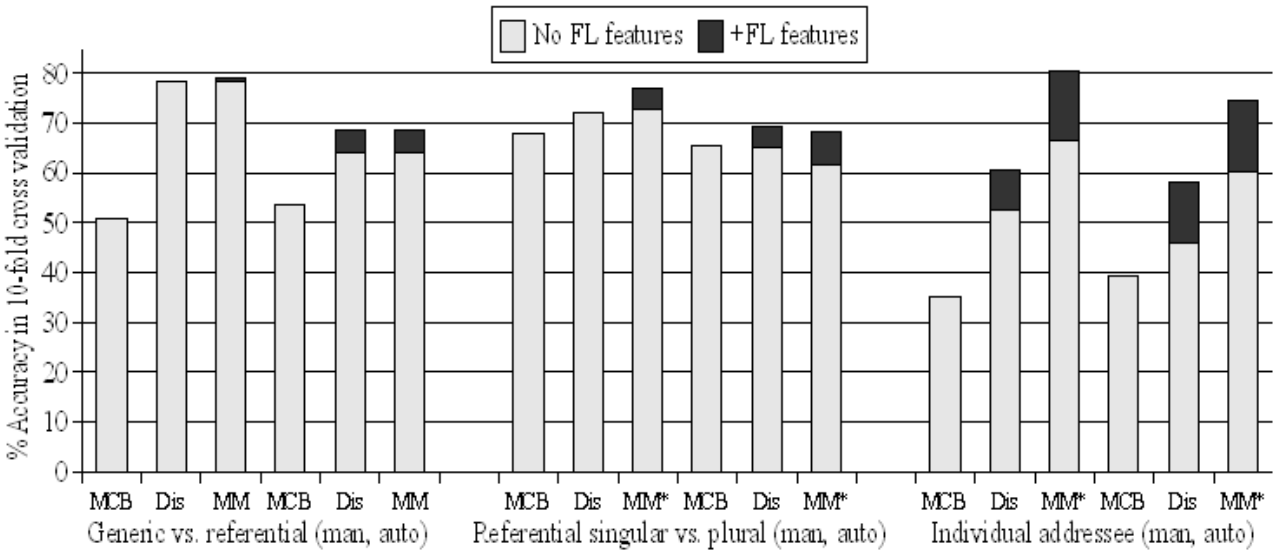


Figure 1: Results for the manual and automatic systems; MCB = majority class baseline, Dis = discourse features, MM = multimodal, * = with Correlation-based Feature Selection (CFS), FL = forward-looking, man = manual, auto = automatic.

ment in the individual addressee resolution task. For the generic *vs.* referential task, the discourse and multimodal classifiers both outperform the majority class baseline ($p < .001$), achieving accuracy scores of 68.71% and 68.48% respectively. In contrast to when using manual transcriptions and annotations (see Section 6.1), removing forward-looking (FL) information reduces performance ($p < .05$). For the referential singular *vs.* plural task, the discourse and multimodal with CFS classifier improve over the majority class baseline ($p < .05$). Multimodal with CFS does not improve over the discourse classifier - indeed without feature selection, the addition of visual features causes a drop in performance ($p < .05$). Here, taking away FL information does not cause a significant reduction in performance. Finally, in the individual addressee resolution task, the discourse, visual (60.78%) and multimodal classifiers all outperform the majority class baseline ($p < .005$, $p < .001$ and $p < .001$ respectively). Here the addition of visual features causes the multimodal classifier to outperform the discourse classifier in raw accuracy by nearly ten percentage points (67.32% *vs.* 58.17%, $p < .05$), and with CFS, the score increases further to 74.51% ($p < .05$). Taking away FL information does cause a significant drop in performance ($p < .05$).

## 8 Conclusions

We have investigated the automatic resolution of the second person English pronoun *you* in multi-party dialogue, using a combination of linguistic and visual features. We conducted a first set of experiments where our features were derived from manual transcriptions and annotations, and then a second set where they were generated by entirely automatic means. To our knowledge, this is the first attempt at tackling this problem using automatically extracted multimodal information.

Our experiments showed that visual information can be highly predictive in resolving the addressee of singular referential uses of *you*. Visual features significantly improved the performance of both our manual and automatic systems, and the latter achieved an encouraging 75% accuracy. We also found that our visual features had predictive power for distinguishing between generic and referential uses of *you*, and between referential singulars and plurals. Indeed, for the latter task, they significantly improved the manual system's performance. The listeners' gaze features were useful here: in our data set it was apparently the case that the speaker would often use the whiteboard/projector screen when addressing the group, thus drawing the listeners' gaze in this direction.

Future work will involve expanding our dataset, and investigating new potentially predictive features. In the slightly longer term, we plan to integrate the resulting system into a meeting assistant whose purpose is to automatically extract useful information from multi-party meetings.

## References


Ron Arstein and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial'06)*, pages 56–63, Potsdam, Germany.

Ilse Bakx, Koen van Turnhout, and Jacques Terken. 2003. Facial orientation during multi-party interaction with information kiosks. In *Proceedings of INTERACT*, Zurich, Switzerland.

Donna Byron. 2004. *Resolving pronominal reference to abstract entities*. Ph.D. thesis, University of Rochester, Department of Computer Science.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky. 2007a. Resolving "you" in multiparty dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September.

Surabhi Gupta, Matthew Purver, and Daniel Jurafsky. 2007b. Disambiguating between generic and referential "you" in dialog. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mark Hall. 2000. *Correlation-based Feature Selection for Machine Learning*. Ph.D. thesis, University of Waikato.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006a. Addressee identification in face-to-face meetings. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL)*, pages 169–176, Trento, Italy.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006b. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40(1):5–23. ISSN=1574-020X.

Natasa Jovanovic. 2007. *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings*. Ph.D. thesis, University of Twente, Enschede, The Netherlands.

Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 144–151, State College, Pennsylvania.

Iain McCowan, Jean Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI Meeting Corpus. In *Proceedings of Measuring Behavior, the 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, Netherlands.

Christoph Müller. 2006. Automatic detection of non-referential *It* in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 49–56, Trento, Italy.

Christoph Müller. 2007. Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 816–823, Prague, Czech Republic.

Andreas Stolcke, Xavier Anguera, Kofi Boakye, Özgür Çetin, Adam Janin, Matthew Magimai-Doss, Chuck Wooters, and Jing Zheng. 2008. The icsi-sri spring 2007 meeting and lecture recognition system. In *Proceedings of CLEAR 2007 and RT2007*. Springer Lecture Notes on Computer Science.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL'03*, pages 168–175.

Yoshinao Takemae, Kazuhiro Otsuka, and Naoki Mukawa. 2004. An analysis of speakers' gaze behaviour for automatic addressee identification in multiparty conversation and its application to video editing. In *Proceedings of IEEE Workshop on Robot and Human Interactive Communication*, pages 581–586.

Koen van Turnhout, Jacques Terken, Ilse Bakx, and Berry Eggen. 2005. Identifying the intended addressee in mixed human-humand and human-computer interaction from non-verbal features. In *Proceedings of ICMI*, Trento, Italy.

Bonnie Webber. 1991. Structure and ostension in the interpretation of discourse deixi. *Language and Cognitive Processes*, 6(2):107–135.