MULTIMODAL ADDRESSEE DETECTION IN MULTIPARTY DIALOGUE SYSTEMS

TJ Tsai[⋆] *Andreas Stolcke*[†] *Malcolm Slaney*[†]

[⋆] University of California Berkeley, Berkeley, CA
[†] Microsoft Research, Mountain View, CA

ABSTRACT

Addressee detection answers the question, “Are you talking to me?” When multiple users interact with a dialogue system, it is important to know when a user is speaking to the computer and when he or she is speaking to another person. We approach this problem from a multimodal perspective, using lexical, acoustic, visual, dialog state, and beam-forming information. Using data from a multiparty dialogue system, we demonstrate the benefit of using multiple modalities over using a single modality. We also assess the relative importance of the various modalities in predicting the addressee. In our experiments, we find that acoustic features are by far the most important, that ASR and system-state information are useful, and that visual and beam-forming features provide little additional benefit. Our study suggests that acoustic, lexical, and system state information are an effective, economical combination of modalities to use in addressee detection.

***Index Terms*—** addressee detection, dialog system, multimodality, multiparty, human-human-computer

1. INTRODUCTION

This paper tackles the problem of addressee detection in multiparty dialogue systems. In multiparty scenarios, users can interact with each other as well as with the system, so it is important for such systems to be able to distinguish between computer-directed and human-directed speech. In order for a dialogue system to operate effectively in open environments, it must be able to determine when a user is speaking to it.

One way to determine whether an utterance is computer-directed is to look at what a person says and how they say it. Shriberg et al. [1][2] explore lexical and prosodic features for addressee detection and find that prosodic features are especially useful. This indicates that the users speak differently depending on who they are talking to. Indeed, several studies show that people tend to speak to the computer more loudly and slowly than they do to humans [3] [4].

Another useful cue is to consider the user’s gaze and head orientation. Skantze and Gustafson [5] demonstrate that head pose can be used to monitor a user’s attention when alternately interacting with a human tutor and a virtual scheduling assistant.

Several works have explored a combination of acoustic, lexical, and visual cues. Bakx et al. [6] use face orientation and utterance length to predict when a user is interacting with a tap-and-talk information kiosk. In a similar study, Turnhout et al. [7] use eye gaze, dialogue state, and utterance length to predict when a user is engaging with an interactive, Wizard-of-Oz (WOZ) information kiosk. Katzenmaier et al. [8] use head pose and simple lexical features to determine when a host is speaking to an imaginary household robot. Baba et al. [3] and Huang et al. [9] use head pose and various prosodic cues to predict the addressee in a multiparty WOZ experiment.

Many of the works above focus on a handful of selected features. The goal of this work is to explore a richer set of features spanning multiple modalities. We do this by using a rich multimodal data set which contains audio, visual, system state, beam-forming, and automatic speech recognition (ASR) information. Vinyals et al. [10] also study addressee detection using this same data set. The focus of this previous work is on exploring discriminative learning techniques that leverage raw streaming features. The focus of our current work is on exploring a rich set of multimodal features and assessing the relative importance of various modalities. In addition, the current work incorporates the best-performing prosodic and lexical features developed in [1], both to validate that methodology on a new corpus and to evaluate other modalities relative to it.

The paper is organized as follows. Section II describes the experimental setup. Section III shares the results of our experiments. Section IV carries out analyses to determine how important each feature modality is. Section V summarizes our findings and concludes the work.

2. EXPERIMENTAL SETUP

We will explain the experimental setup in four parts: the data, the features, the classifier, and the evaluation method.

2.1. Data

We used data from a multiparty dialog setup described by Bohus and Horvitz [11]. The scenario involves groups of 2 and 3 people playing a trivia question game with a computer agent. The computer agent is a talking face on a computer monitor equipped with a 4 element linear microphone array and a wide-angle camera for visual tracking. The agent asks the group questions, confirms what one participant says with one other participant, and then tells them if their answer is correct.

The data includes audio, video, beam-forming, system state, and ASR information. The beamformed audio was automatically segmented by a speech activity detector, and the resulting utterances were annotated with speech, speaker, and addressee information. We considered an utterance to be directed towards the computer if any speech within the utterance is addressed to the computer.

We divided the utterances into 15 folds based on the groupings of participants. Eight of the folds were used for training and the other seven for testing. The training and testing sets had 2001 and 1952 utterances, respectively.

The first author performed the work as an intern at Microsoft Research. Research conducted while the second and third authors were with Microsoft Research. The third author is now working at Google.

2.2. Features

We explored 5 different feature modalities: acoustic, visual, system, beam-forming, and ASR. We describe the features in each modality below.

Acoustic. We extracted three families of acoustic features. The first family of acoustic features are energy features. These features are measures of frame-level energy over various intervals of time. These intervals include frames up to three seconds before and after the utterance. The intuition behind these features is that people tend to speak more loudly when addressing the computer, so energy measures may help discriminate between computer- and human-directed utterances.

The second family of acoustic features are energy change features. These features compute the difference in energy between two adjacent intervals in time. Again, the intervals span up to 3 seconds before and after the utterance. The intuition here is that people tend to pause after speaking to the computer, waiting for the computer's response. Energy change features can simultaneously capture the elevated volume during the utterance and the pause immediately afterwards in a computer-directed utterance.

The third family of acoustic features characterize the temporal shape of the speech energy contour, as first described by Shriberg et al. [1]. We compute zeroth and first-order mel-frequency cepstral coefficients (MFCCs) every 10 milliseconds, and we characterize the contours of these values over windows of 200 milliseconds by computing a discrete cosine transform (DCT) in the temporal domain. We retain the first 5 DCT values for c0 and the first 2 DCT values for c1, resulting in a 7-dimensional feature vector for every 200ms window. Note that the other two families of acoustic features are utterance-level features, while we compute energy contours at the frame level. We trained two Gaussian mixture models (GMMs): one GMM to model energy contours in human-directed speech and one GMM to model energy contours in computer-directed speech. The *utterance*-level energy contour features are the log likelihood ratios of these two GMMs.

Visual. We extracted three families of visual features. The first family of visual features are measures of the amount of movement. The idea here is that people tend to be more stationary when interacting with the computer than with other people. Some examples of these features include: (1) the variance of the speaker's face pose angle and (2) the average variance of all the participants' face locations. We computed these measures over various intervals up to 3 seconds before and after the utterance.

The second family of visual features are measures of face orientation. Where a person is looking can be a useful indicator of who they are talking to. Since eye gaze information was not available, we use face-pose angles instead. Some examples of the face orientation features include: (1) the speaker's average pose angle in the up/down direction, (2) the speaker's average pose angle away from the computer in the left/right direction, and (3) the fraction of speaker's pose angle estimates that were unavailable. The third example refers to the fact that face pose estimates could not be computed when a person turns their face too far to the side. The fraction of pose-angle estimates that could not be computed can thus still be a useful indication of face orientation. We computed these measures over various intervals in time to account for lags between when speech begins and when the face turns.

The third family of visual features are measures of physical distance between the participants. The idea here is that the distance between two people may be a social signal indicating how comfortable they feel with each other. Two people who feel uncomfortable around each other will probably stand further apart and are less likely to have discussions together. We used pixel distances between participants' face locations as a proxy for physical distance. Some examples of these features include: (1) the distance between the speaker and the nearest actor, and (2) the change in distance between the speaker and farthest actor over two neighboring intervals of time. To compute a single distance metric over an interval of time, we considered the minimum, mean, and maximum of constituent frame-level distances. As before, we computed these measures over various intervals of time.

For more detailed information on how the system did face detection and pose estimation, see the earlier work by Bohus and Horvitz [12] and corresponding references.

System. The system features are various indicators of the system state. The idea here is that the context in which a person speaks is predictive of his or her linguistic behavior. Some examples of these features include: (1) the number of participants in the interaction, (2) the time elapsed since the computer agent last spoke, and (3) the dialog act type of the last computer agent utterance (question, confirmation, answer, etc). Note that, unlike the previous features, some of the system features are categorical, not numerical, in nature.

Beam-forming. The beam-forming features describe the distribution of beam values, which are indicative of the direction of incoming audio. A wide spread of beam values suggests that multiple people are talking. In this way, the distribution of beam values can be an indicator of the level of discussion or activity among the participants. Some examples of these features include: (1) the variance of beam values, (2) the range of beam values (i.e. the difference between maximum and minimum), and (3) the fraction of beam values falling within a certain range. As before, we computed these measures over various intervals of time.

ASR. We extracted two families of ASR features. The first family of ASR features model lexical n-grams as described by Shriberg et al. [1][2]. We trained two maximum-entropy trigram language models: one model for human-directed utterances and one model for computer-directed utterances. The utterance-level feature is the log likelihood ratio from these two models. The intuition here is that people tend to use different words, phrases, and syntactic patterns as a function of who they are addressing.

The second family of ASR features describe various properties of the hypotheses generated by the speech recognition engine. These include features such as: (1) the duration of the utterance, (2) the confidence of the top hypothesis, (3) the number of hypotheses, and (4) the number of words in the hypotheses. Classifying utterances based on ASR confidence capitalizes on the fact that human-directed speech tends to be less well-matched to the recognizer's acoustic and language models than computer-directed utterances.

Feature Summary. In total, we extracted 117 different features. Table 1 shows a breakdown of the feature count by group.

We took special care to avoid extracting model-based features (lexical n-gram and energy contour features) using models that were trained on the same data for which we are computing features, as this would lead to features that are optimistically biased. For feature computation on the training set, we trained feature models (language models and GMMs, respectively) on 7 training folds at a time, then computing features for the 8th fold, and did so round-robin for all training folds. For the test set, the corresponding features were then computed using models trained on the entire 8-fold training set.

Feature Type	Count
Acoustic	47
Visual	41
System	6
Beam-forming	16
ASR	7
Total	117

Table 1. Breakdown of feature count by modality.

2.3. Classifier

We used adaboost with tree stumps as our classifier. We did experiment with several other classifiers, and we found that adaboost performed the best and is broadly representative of results with different feature subsets. We selected the number of trees based on cross-validation experiments and describe the classifier performance in the next section.

2.4. Evaluation Method

We evaluate models, features, and modalities using methodology commonly used for detection tasks. A good way to visualize the performance of a detection system is to plot its detection error tradeoff (DET) curve [13]. The DET curve shows the tradeoff between the false alarm rate (on the x-axis) and the missed detection rate (on the y-axis). Both axes use a normal deviate scale to achieve a roughly linear plot shape. Sometimes it is more convenient to express system performance in a single number, especially when comparing several systems and when the DET curves run roughly in parallel. In that case we use equal error rate (EER), which refers to the point on the DET curve where the false alarm and missed detection rates are equal. Importantly, the EER is invariant to changing class priors, and also equals the overall classification error rate at the corresponding operating point. Also note that a system outputting random decisions would have an EER of 50%.

3. RESULTS

The DET curves in Figure 1 show the performance with all modalities combined ("top 5 modalities"), the single best modality ("top 1 modalities"), as well as intermediate systems obtained by adding additional modalities in order of their individual performance. The order in which the modalities are added is acoustic (most important), ASR, system, visual, and beam-forming. How this order was determined will be discussed in section IV. This figure tells a system designer how much is gained at each step by implementing the next most important feature modality.

There are three things to notice about Figure 1. First, there is significant improvement in including multiple modalities. When we compare the performance of the system using one modality and the system using all 5 modalities, the equal error rate (EER) falls from 13.9% to 9.8%. Second, adding more modalities yields increasingly marginal gains. We can see that each new modality we add helps less and less. Beyond the top 3 modalities (acoustic, ASR, system), the gains are minimal. In this case, it may not be worth the effort to compute visual features for such marginal gains. We often see the law of diminishing returns when combining features and combining systems, and here we see the law of diminishing returns with combining modalities as well. Third, the performance of the system with the single best modality (acoustic) has very poor performance in the low false alarm region. Note that all five DET curves have roughly converged in performance for low miss rates ($< 5\%$), but that the system with only one modality has much higher miss rate for low false alarms. Here, we are detecting computer-directed utterances, so a low false alarm rate means that we want to keep human-directed speech from the system. In these scenarios, including two or more modalities will significantly improve system performance.

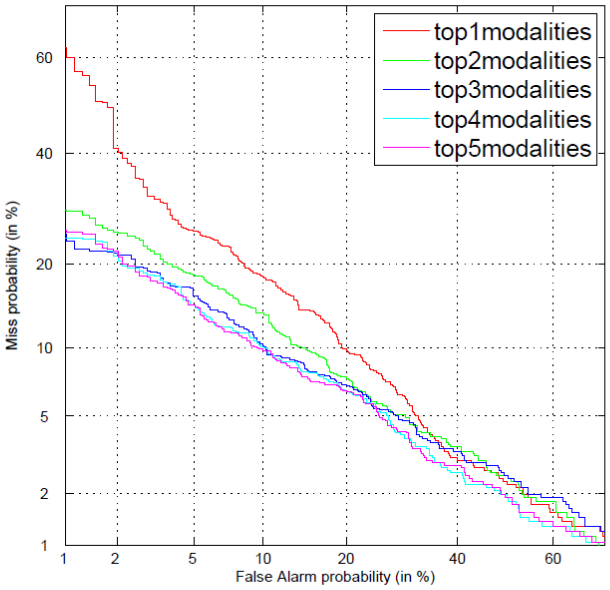

Fig. 1. DET curves showing the incremental improvement by modality. The order of modalities is acoustic (most important), ASR, system, visual, and beam-forming (least important). Each curve shows the performance when the top N feature modalities are used.

4. ANALYSIS

In this section we assess the relative importance of the various feature modalities.

One way to determine the importance of different feature modalities is to look at the relative influence of individual features [14][15]. Relative influence is the reduction in loss function attributable to a single feature, normalized by the total loss reduction by all features. It is a measure of how much an individual feature influences the adaboost prediction. So, a feature with 0% relative influence does not affect the ensemble prediction at all, while a feature with 100% relative influence would deterministically control the prediction.

Figure 2 shows the relative influence of all 117 features in our adaboost model, grouped by modality. Within each grouping, the features are sorted in decreasing order of relative influence. A brief glance at this figure immediately reveals the major trends among the various modalities: The top several acoustic features dominate. Beam-forming features don't matter. The top few ASR and system features are useful. The rest don't seem to contribute much. Note that this plot suggests the ordering of importance used in figure 1.

Another way to assess the importance of different modalities is to run full end-to-end experiments using one group of features at a time. These experiments will reveal how well we can do addressee detection when only using information from a single modality. The

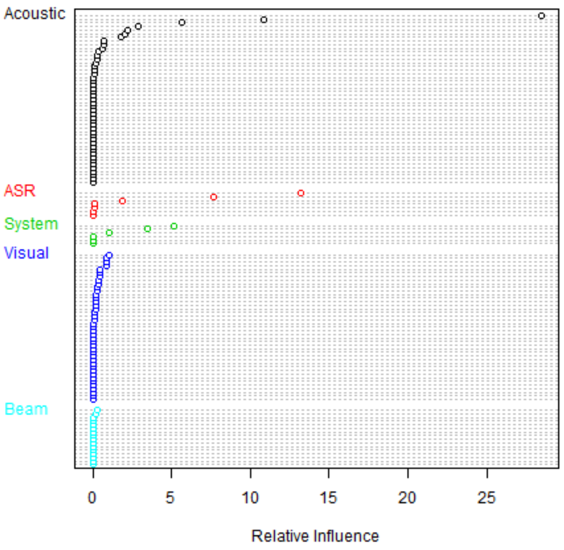

Fig. 2. Relative influence of all 117 features in the adaboost model. The features are grouped first by modality, then in decreasing order of influence.

Feature Modality	Leave-In EER	Leave-Out EER
Acoustic	13.88%	16.39%
ASR	19.01%	11.33%
System	30.02%	10.60%
Visual	32.68%	9.99%
Beam	40.98%	9.94%
All	9.84 %	-

Table 2. Equal error rate of adaboost classifier when only using a single feature modality (middle column) and when leaving out a single feature modality (right column). The performance with all feature modalities is shown on the bottom row for reference.

middle column of Table 2 shows the EERs for our leave-one-group-in experiments. Similarly, the rightmost column of Table 2 shows the EERs when we *remove* one feature modality at a time. So, for example, the system with all modalities *except* acoustic features has an EER of 16.39%. Note that for our leave-one-group-out experiments, a *higher* EER suggests that the excluded modality is more important. We see that both the leave-one-in and leave-one-out experiments suggest the same ordering of importance used in figure 1.

The fact that visual features were not very important may seem surprising, especially since visual information such as gaze and face orientation are important cues in human interactions. The main reason that face orientation was not very important is because the computer agent was a major situational attractor: people continued looking at the screen even as they talked amongst themselves. This phenomena has been described by social psychologists [16], and has also been observed in several other related studies [6][7][8][9]. Because people's default face orientation for this task was towards the computer, face pose provides little useful information.

5. CONCLUSION

We have proposed a multimodal addressee-detection system and assessed the relative importance of various modalities in predicting addressee. In a multiparty human-computer dialogue scenario, we find that acoustic information is most useful, dominating the other modalities in importance. Lexical and dialog state information are also useful, providing significant performance gains. Visual and beam-forming information provide little additional benefit in our experiments. Our results suggest that audio information (both prosodic and lexical) and system state information are a good combination of modalities to use, providing a good balance between performance and economy of implementation.

6. ACKNOWLEDGMENTS

We greatly appreciate help from Dan Bohus and Oriol Vinyals in providing the data, answering questions, and providing useful feedback.