A Study of Multimodal Addressee Detection in Human-Human-Computer Interaction

T. J. Tsai, *Student Member, IEEE*, Andreas Stolcke, *Fellow, IEEE*, and Malcolm Slaney, *Fellow, IEEE*

***Abstract*—The goal of addressee detection is to answer the question, "Are you talking to me?" When a dialogue system interacts with multiple users, it is crucial to detect when a user is speaking to the system as opposed to another person. We study this problem in a multimodal scenario, using lexical, acoustic, visual, dialogue state, and beamforming information. Using data from a multiparty dialogue system, we quantify the benefits of using multiple modalities over using a single modality. We also assess the relative importance of the various modalities, as well as of key individual features, in estimating the addressee. We find that energy-based acoustic features are by far the most important, that information from speech recognition and system state is useful as well, and that visual and beamforming features provide little additional benefit. While we find that head pose is affected by whom the speaker is addressing, it yields little nonredundant information due to the system acting as a situational attractor. Our findings would be relevant to multiparty, open-world dialogue systems in which the agent plays an active, conversational role, such as an interactive assistant deployed in a public, open space. For these scenarios, our study suggests that acoustic, lexical, and system-state information is an effective and practical combination of modalities to use for addressee detection. We also consider how our analyses might be affected by the ongoing development of more realistic, natural dialogue systems.**

***Index Terms*—Addressee detection, beamforming, dialogue system, head pose, human-human-computer, multimodal, multiparty, prosody, speech recognition.**

I. Introduction

MORE and more, speech-enabled dialogue systems are embedded in our environment in entertainment systems, mobile phones, and wearable accessories. These devices increasingly employ multimodal sensors and allow for natural interactions via conversational software agents. The confluence of these trends exacerbates the problem of knowing when to interpret the user's inputs as system-directed, rather than as unrelated actions or communication with humans. For vision-based systems, this is often called the "Midas Touch" problem. For speech-based systems, this problem is known as addressee detection (AD).

Addressee detection tries to answer the question, "Who are you talking to?" In human-to-computer (H-C) interactions, the problem includes the rejection of self-talk and background speech, but becomes harder in a multi-human-computer scenario, since users now have a choice of talking to the system as well to humans. Traditionally, user interfaces have been engineered to remove addressee ambiguity (e.g., through prompted interaction or push-to-talk), or to assume that all potential inputs are system-directed and to reject them based on failure-to-recognize or failure-to-interpret [1], [2]. Both approaches are no longer feasible as systems allow natural interactions with essentially unlimited domain coverage (e.g., the input could comprise a general search query in conversational form). We must therefore look to more comprehensive cues and more sophisticated classification methods to determine addressee for a potential input.

The work reported here improves upon two previous lines of work on human-human-computer (H-H-C) addressee detection. One is our previous work on multimodal interfaces and the exploitation of multiple modalities for addressee classification [3], [4]. The other is the characterization of speaking style and lexical content, which in conjunction yield highly accurate addressee estimates based on speech alone [5]–[7]. The present paper has two main contributions. The first main contribution is to present the most comprehensive multimodal approach to date, which includes a much wider range of information than previous approaches and incorporates recent advances in speech-only AD. Specifically, we combine prosodic, lexical, visual, dialogue state, and beamforming information as can now be obtained from consumer-grade sensors and speech technology. The second main contribution is to determine what type of information is most useful for the AD task in the multimodal scenario, both at the feature level and by modalities in aggregate. This type of analysis can both guide future research into the problem and inform engineering solutions that need to achieve best possible results with the least resources and complexity.

The paper is organized as follows. Section II reviews related work. Section III describes the experimental setup, including a detailed explanation of our multimodal addressee detection system. Section IV reports the results of our experiments, while Section V describes various analyses to determine detailed contributions to overall performance. Finally, Section VI summarizes the findings and open questions.

Manuscript received February 03, 2015; revised May 22, 2015 and July 02, 2015; accepted July 06, 2015. Date of publication July 09, 2015; date of current version August 10, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sheng-Wei Chen.

T. J. Tsai was with Microsoft Research, Mountain View, CA 94043 USA. He is now with the University of California at Berkeley, Berkeley, CA 94704 USA (e-mail: tjtsai@icsi.berkeley.edu)

A. Stolcke is with Microsoft Research, Mountain View, CA 94043 USA (e-mail: stolcke@icsi.berkeley.edu).

M. Slaney was with Microsoft Research, Mountain View, CA 94043 USA. He is now with the Machine Hearing Group, Google Research, Mountain View, CA 94043 USA (e-mail: malcolm@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2454332

II. Related Work

In this section we summarize previous work and explain how this paper fits into the broader landscape of research.

A natural starting point in tackling addressee detection is understanding human behavior and language in group situations involving a computer or robot. There is a large body of work in the human computer interaction field on this topic, from which we will sift out three key observations. The first observation is that people tend to look at the device they are talking to. Works by Brumitt and Cadiz [8] and Maglio *et al.* [9] show that, when confronted with multiple speech-enabled devices, people specify the recipient of their request through eye gaze. The second observation is that how a person sees a robot strongly shapes how he or she interacts with it. For example, Lee *et al.* [10] demonstrate that people interact with a robot differently depending on whether they see the robot as an information kiosk or a receptionist. The third observation is that humans tend to rely on several different types of information to discern the addressee in H-H-C interactions. Several previous works have examined the various cues that human evaluators use to discern the intended addressee in recordings of multiparty interactions involving a computer or information retrieval agent [11]–[13]. These studies find that humans use a combination of lexical, gaze, and prosodic information. All of the above work motivates our multimodal approach to AD.

Next, we look at previous work in automatic addressee detection in H-H-C scenarios. As mentioned above, the way people interact with a computer system depends heavily on the nature of the computer system. For this reason, we will distinguish between two separate categories: scenarios in which the computer system is passive and scenarios in which the computer system is an active, conversational agent.

A number of works have focused on H-H-C scenarios where the computer system is passive and simply receives commands. Shriberg *et al.* [5], [7] focus on audio information only and explore various lexical and prosodic features for addressee detection. The setup involves two users trying to accomplish a web-browsing task using speech commands to control the system. These studies show that acoustic-prosodic features modeling energy contour and raised voice are very effective, suggesting that speakers use different speaking styles depending on who they are talking to. Bakx *et al.* [14] explore face orientation and utterance length to do addressee detection. In this scenario, a user and partner use a tap-and-talk information kiosk to buy train tickets. Katzenmaier *et al.* [15] explore head pose and lexical features based on automatic speech recognition (ASR) hypotheses. In this setup, a host introduces an imaginary household robot to a guest and demonstrates some of its functionality. The studies by Bakx and Katzenmaier both find that the computer or robot is a major situational attractor. In other words, people continued looking at the computer while talking to each other. Note that the above works consider a variety of features for AD, but each work only considers a few selected features or modalities.

Finally, we consider works that investigate H-H-C scenarios where the computer is an active, conversational agent, which is the scenario of focus in this study. In these scenarios, the computer both listens and speaks during interactions with users. Baba, Huang, and colleagues [16], [17] explore prosodic and head pose information to predict the addressee in a multiparty, Wizard-of-Oz (WOZ) experiment. They find that intonation, volume, and rate of speech are useful features, but that head direction alone is insufficient to make good predictions. Van Turnhout *et al.* [18] explore eye gaze, dialogue state, and utterance length as predictors for addressee detection. This is also a WOZ setup in which two users engage with an interactive information kiosk to book train tickets. They also find that the screen is a major situational attractor. Skantze and Gustafson [19] use head pose to monitor a user's attention when he or she alternately interacts with a human tutor and an interactive virtual scheduling assistant. This study finds that head pose is an effective cue for predicting addressee.

One shortcoming of all of the above approaches is that they generally focus on only a few selected features or modalities, which makes it difficult to perform a systematic study of the importance of various features or modalities. Understanding which individual *features* are most important is useful in guiding the development of more effective features for AD. Understanding which *modalities* are most important is useful for economy of implementation, especially since adding a modality often incurs a significant cost in equipment (e.g., adding a microphone array or video camera), in system complexity (i.e., having to incorporate multiple modalities of information), and in data processing (e.g., running face detection on a video stream). Such understanding will be of practical use to researchers building a system that requires addressee detection.

The goal of this current work is to undertake such a systematic study: we adopt a comprehensive multimodal approach to AD in a multiparty dialogue system in which the agent is an active, conversational agent. We are aware of only one other work, by Vinyals *et al.* [4], which examines AD in the context of a rich multimodal data set. Our current work uses the same rich multimodal data set as their study, but with a different focus. The Vinyals *et al.* study explores discriminative learning techniques applied to raw data streams, whereas our focus is on exploring a much richer, broader set of multimodal features in order to facilitate the study of feature importance as described above. These multimodal features cover a wider range of information than any of the individual works previously described, and they incorporate recent advances in strong audio-only (prosodic and lexical) features [5]. Furthermore, this work presents the first detailed and comprehensive analysis of the relative importance of and synergies among different features and modalities for the AD task.[1]

III. Experimental Setup

We will explain the experimental setup in three parts: the data, the features, and the classifiers.

A. Data

In order to explore the use of multimodal features for AD in H-H-C interactions, we need a data set that satisfies several

[1]This paper extends our earlier work [20], expanding the experiments to include several different classifiers, investigating the most important individual features, studying the effect of absolute energy and post-utterance context, looking more closely into why the visual modality was not very important, and conducting statistical tests on differences in equal error rate.

Fig. 1. Snapshot of the collection setup.

key criteria. First, the data must contain multimodal information. Some previous work has focused on addressee detection using only audio information, so the data sets in those experiments may not be suitable for our purposes. Second, the data should comprise multiparty interactions in which two or more humans simultaneously interact with a dialogue system. Third, the interactions should allow people the option to talk to the computer or to another person. If people are only allowed to interact with the computer, the problem of AD becomes trivial. Fourth, the data must have ground truth annotations of whom is being spoken to. Annotating addressee information is time-consuming and is perhaps the most restrictive of the criteria.

Based on these criteria, we selected a data set from a multiparty dialogue setup described by Bohus and Horvitz [21]. The scenario involves groups of two or three people playing a trivia question game with a computer agent. The computer is represented as a talking face displayed on a 19-inch computer monitor. The agent has controllable head poses and limited facial gestures, and it engages with the participants through dialogue and face movement. 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 set was designed to study computational models of multiparty turn-taking, and it encourages natural, fluid interactions. Fig. 1 shows a snapshot of the data collection setup.

Participants were recruited in pairs of people who knew each other. They were divided into 15 groups of 4, where each group consists of two pairs. Within each group of 4, every possible subgroup of two or three was formed, and each subgroup played one game together. This results in 10 games per group and 150 games in total.

The data available for our use included audio, video, beamforming, system state, and ASR information. The audio was recorded by a linear microphone array, which can be seen in Fig. 1 as a thin rectangular bar located directly above the upper bezel of the monitor. The array is symmetric and contains four uni-directional microphones, where the outer pair of microphones are located 190 mm apart and the inner pair are located 55 mm apart. The audio is processed with the built-in Windows microphone array support, which provides acoustic echo cancellation, minimum variance distortionless response beamforming, and source localization in 10 degree increments. The audio is further processed with the integrated Windows speech recognizer using simple grammars. The video data was collected with a wide-angle AXIS 212 camera with a resolution of 640×480 pixels. The camera can be seen in Fig. 1 located above the microphone array. The system processes the video data in real-time to track the faces of each participant. It performs face detection on each frame, and then associates detections across frames using a proximity based algorithm [22]. For each detected face, the system runs a Bayesian pose-tracking algorithm [23] that produces estimates of 3-D head orientation. In addition to the audio and video information, the system logs information describing various aspects of the interaction, such as how many participants there are, what the computer agent is saying, and who the agent is looking at. More detailed information about the setup can be found in the original works by Bohus and Horvitz [24], [21]. Note that all of the information described above is collected or computed in real-time during the actual interaction, but we perform our AD experiments in an offline setting.

In addition to the raw data, the data set includes manual annotations. The audio was automatically segmented by a speech activity detector, and the resulting utterances were manually annotated with speech, speaker, and addressee information. Because the interactions between participants are unscripted, overlapped speech is a common occurrence. When there is overlapping speech, the speaker, speech, and addressee information was annotated for each stream of speech separately. Note that when two people are speaking at the same time, one might be talking to the computer while the other is talking to another person. To handle cases like these, we considered an utterance to be directed toward the computer if any speech within the utterance is addressed to the computer.

Though this data set satisfies the essential criteria mentioned above, it is important to point out that there are virtual agents with more sophisticated capabilities than the computer agent in our study. For example, DeVault *et al.* [25] have developed a full-body virtual human interviewer that tracks the interviewee's face pose and location, gaze direction, and facial expression. This additional information allows the agent to respond in a more sensitive and personal way, which creates a more natural social interaction. Bohus *et al.* [26] have also deployed physical robots in public spaces that engage with one or more people, engaging in dialogue and giving directions both verbally and with physical gestures. When interpreting the conclusions and findings of our study, it will thus be important to keep in mind the limitations of the computer agent and the ongoing development of more natural dialogue systems.

B. Data Usage

For our experiments, we divided the utterances into 15 folds, which correspond to the 15 groups of 4 participants. We used 8 of the folds for training and the other 7 for testing. The training and test sets had 2001 and 1952 utterances, respectively. Some features are computed as log likelihood ratios of class-specific models, and require training data for those models.[2] For these likelihood-ratio features, we were careful to avoid reuse of the data that could bias the features. When computing likelihood-ratio features on the training set, we train the utterance class

[2]These are log likelihood ratios that aggregate a variable number of samples at the utterance level, and include the lexical n-gram and energy contour features described in the next subsection.

models on 7 training folds and used the resulting models to compute features for the 8th fold. We repeated this in a round-robin fashion to compute features on all 8 training folds. When computing likelihood-ratio features on the testing set, we used models trained on all 8 training folds. This arrangement ensures that our model-based features are not optimistically biased.

C. Features

We explored five different modalities of features: acoustic, visual, system, beamforming, and ASR. For each modality, we describe the features extracted and the intuition behind their design. The number of features is shown in parentheses.

Acoustic: We extracted three families of acoustic features. The first family consisted of energy features, i.e., measures of frame-level energy over various intervals of time (21 features). Examples include: (1) the average energy during the first third of the utterance; (2) the maximum frame-level energy throughout the utterance; and (3) the average energy during the 1-second interval preceding the utterance. The various intervals include frames up to 3 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 consisted of energy change features (24 features). These features compute the difference in energy between two neighboring intervals in time. Examples include: (1) the difference between the average energy during the utterance and the average energy during the 2-second interval after the utterance, and (2) the difference between the maximum frame-level energy during the first third of the utterance and the maximum frame energy during the 1 second before the utterance. 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 while 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 (2 features), as first described by Shriberg *et al.* [5]. Zeroth and first-order mel frequency cepstral coefficients are computed every 10 milliseconds, and the contours of these values over windows of 200 milliseconds are characterized by computing a discrete cosine transform (DCT) in the temporal domain. The first 5 DCT values for cepstral coefficient c0 are retained, as are the first 2 DCT values for c1, resulting in a 7-dimensional feature vector for every 200 ms window. Shriberg *et al.* observed that users employ a more regular rhythmic up-and-down energy pattern when talking to the computer versus to humans, similar to how one might talk to a child or linguistically handicapped person, and this difference in contour shape is captured by the resulting distributions of DCT values. Note that the other two families of acoustic features are utterance-level features, while the energy contours above are computed at a frame level. To arrive at utterance-level features we train two Gaussian mixture models (GMMs): one to model the feature vectors in human-directed utterances and one for computer-directed utterances. The log likelihood ratio computed from these two class-conditioned models becomes a single *utterance*-level energy contour feature value, and is used alongside the other utterance-level features. An alternate version of this feature normalizes the log likelihood by utterance length, i.e., by the number of frames.

Visual: We extracted three families of visual features. The first of these is designed to measure the amount of movement (12 features). The idea here is that people tend to be more stationary when interacting with the computer than with other people. Examples of these features include: (1) the variance of the speaker's face location; (2) the variance of the speaker's face pose angle; and (3) 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. When we computed features like the variance of the speaker's face location, we used ground truth annotations of whom the speaker is, rather than the system's estimates. By removing the uncertainty of the system's estimate, we can more clearly discern how important this type of information is, independent of how robust the speaker identification estimate is. If the speaker-specific features turn out to be very important, we should interpret performance numbers as an upper bound, assuming perfect speaker identification. If the speaker-specific features turn out not to be important, we can more confidently conclude that this type of information is not very useful for the given task.

The second family of visual features is designed to capture face orientation (11 features). Based on the research reviewed earlier, a person's gaze can be a useful indicator of who they are talking to. We would have liked to use eye-gaze information in our study, but this was (and still is) difficult to obtain at the distances involved. Common eye trackers using a single camera without mechanical tracking might only be able to analyze the eyes out to 90 cm [27], which is sufficient for a seated desktop interface but not nearly far enough for standing interactions. Head pose information is not as accurate an indication of people's attention as eye gaze, but still has the potential to inform an AD system.

Examples of 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 normalized pose angle in the second example is a measure that removes the effect of speaker location. So, regardless of whether the speaker is standing on the left or right, the normalized angle simply measure the angle away from the computer. The third example refers to the fact that face pose estimates cannot 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 compute 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 (18 features). 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

farther apart and will be less likely to have discussions together. Because depth estimates were not available, we used pixel distances between participants' face locations as a proxy. Some examples of these features include: (1) the distance between the speaker and the nearest/farthest actor, and (2) the change in distance between the speaker and nearest/farthest actor over two neighboring time intervals. To compute a single distance metric over an interval of time, we considered the minimum, mean, and maximum of constituent frame-level distance values. 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 [24] and corresponding references.

System: System features comprise various indicators of the system state, including the state of the dialogue manager (6 features). The idea here is that the context in which a person speaks is predictive of his or her lingustic behavior. Some examples of these features include: (1) the number of participants in the interaction; (2) the time elapsed since computer agent last spoke; and (3) the dialogue act type of the last computer agent utterance (question, confirmation, answer, etc.). Note that, unlike most of the features described earlier, several of the system features are categorical, rather than numerical, in nature.

Beamforming: The beamforming features include various descriptors of the distribution of beam values, which indicate the direction of incoming audio (16 features). 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. 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. Again, we computed these measures over various time intervals.

ASR: We extracted two families of ASR features. The first of these model lexical n-grams (2 features) in the same way as described by Shriberg *et al.* [5], [7]. We trained two maximum-entropy trigram language models: one model for computer-directed utterances and another for human-directed utterances. Similar to the energy-contour features, we computed the log likelihood ratio from these two models to get a single utterance-level feature value. An alternate version of this feature normalizes the log likelihood by the number of words in the utterance. The intuition behind the n-gram modeling is that people tend to use different words, phrases, and syntactic patterns depending on who they are addressing.

The second family of ASR features describes various properties of the hypotheses generated by the speech recognition engine (5 features). These include: (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 top (or all) 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 I shows a breakdown of the feature count by modality.

TABLE I
BREAKDOWN OF FEATURE COUNT BY MODALITY

Feature Type	Count
Acoustic	47
Visual	41
System	6
Beamforming	16
ASR	7
Total	117

Computational Complexity: A meaningful discussion of computational complexity must discuss the *marginal* cost of computing AD features. AD is not very useful in isolation, but is used in conjunction with other system components. It can thus reuse a lot of the work that is necessary for other system components. For example, a dialogue system by its very nature must compute an auditory spectrogram and perform ASR. So, the marginal cost of the acoustic AD features consists of (1) computing average energy across various intervals of time, which can be done very efficiently using an integral representation of frame energy (e.g. [28]), and (2) computing two discrete cosine transforms every 200 ms for the energy contour features. The marginal cost of the lexical AD features is computing the log likelihood ratio between two language models for the hypothesized word sequence. The marginal cost of the visual AD features is computing simple statistics such as average or variance across a set of face locations, head pose estimates, or distances between faces. Note that in our experimental setup, most of the heavy visual processing (i.e. face detection and pose estimation) has already been done by the dialogue system, though an audio-only dialogue system would incur a very heavy marginal cost in computing visual features for AD. The marginal cost for beamforming features is likewise computing simple statistics across the beam estimates in each utterance. The marginal costs for other features not explicitly mentioned above are trivial, such as accessing system state information or using the confidence of the top ASR hypothesis.

Implementation: The ASR, face detection, and face pose estimates were done in real-time during the actual interactions. We then extracted the necessary information from log files, carried out the "marginal" processing described above, and conducted our AD experiments in an offline manner. Even though our implementation was entirely offline, it is useful to point out that the marginal costs described above are relatively small, and could easily be performed in real-time. However, note that some features incorporate context outside of the duration of the utterance itself, including up to 3 seconds *after* the utterance ends. These features would require a delay in processing, and would probably not be acceptable for use in a real-time system. However, it will be useful to know if the context *after* an utterance ends provides useful information in predicting addressee. We will investigate this in Section V-C.

D. Fusion

There are many different methods of combining information from multiple modalities. As Atrey *et al.* point out in their survey of multimodal fusion strategies [29], two key choices

are the *level* of fusion and the *method* of fusion.[3] The level of fusion can be at the feature level (early fusion), the decision level (late fusion), or a combination or blend of both. The method of fusion can be rule-based, classification-based, or estimation-based. Much of the recent work in classification-based fusion methods have explored fusing representations within a deep belief network or deep Boltzmann machine [31]–[34], or combining the benefits of generative and discriminative models in a hybrid architecture [35].

In this work, we adopt a similar approach to Perez-Rosas [36], in that we perform early fusion and compare the performance of various subsets of features. We have selected this methodology for two main reasons. First, because of the limited amount of labeled training data, we prefer a simple architecture so as not to introduce too many parameters. Second, because one of our main goals is to understand and analyze the relative importance of various features, we prefer an architecture that facilitates a straightforward analysis of feature importance. For these two reasons, we adopt a simple early-fusion methodology.

E. Classifiers

We experimented with four different types of classifiers: logistic regression, decision tree, random forest, and Adaboost with tree stumps. Logistic regression models the probability of a binary outcome as a function of the features using a linear logistic function [37]. Decision trees recursively partition the data to minimize some measure of node impurity, and they assign each partition a categorical or numerical prediction [38]. Random forest is an ensemble-learning method, which differs from decision trees in two ways: (1) it trains multiple decision trees, each based on a bootstrap sample of the data, and (2) each decision tree is grown using a modified tree learning algorithm that selects a random subset of features at each candidate split [39]. Adaboost is a meta-algorithm for combining the predictions of a set of weak learners, where each subsequent weak learner is selected to address the mistakes of previous classifiers [40].

We experimented with both classification trees and regression trees, and we found that regression trees have slightly better performance, both for the single tree classifier and the random forest model. We report only the results with regression trees in this work. We also tried Adaboost with trees of greater depth but found the results to be no better. We only report the results with tree stumps. For the random forest and Adaboost models, we selected the number of trees to ensure convergence. We will investigate the performance of these classifiers in the next section.

F. Evaluation

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 [41]. 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. Statistical significance of EER differences is assessed using a McNemar (matched pairs) test, as described in [42] (see Section II-C4). Also note that a system outputting random decisions would have an EER of 50%.

[3]This provides one perspective on multimodal fusion strategies. The survey paper by Lalanne *et al.* [30] provides another perspective.

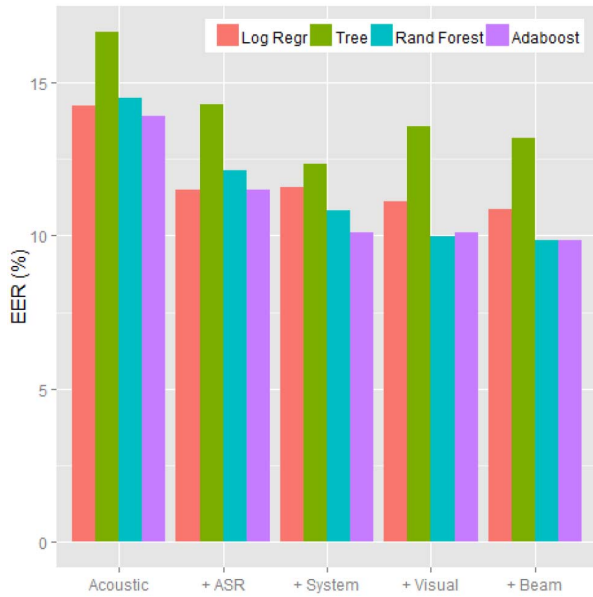

Fig. 2. Equal error rates of all four classifiers when adding more and more feature modalities. The leftmost column shows the performance with only the most important feature modality (acoustic), and the rightmost column shows the performance with all five feature modalities.

IV. Results

In this section we report overall performance of our multimodal system for the different classifiers, as well as performance of subsets of the modalities.

A. Performance With All Classifiers

First we show the overall results of all four classifiers and the benefit of using multiple modalities. Fig. 2 shows the EER for the four classifiers when they are incrementally provided with more and more modalities, where the modalities are added in order of their individual performance. So, for example, the leftmost group of bars shows the performance when only the acoustic modality is available, and the rightmost group of bars show the performance of the classifiers when all five modalities are used. This plot shows a system designer how much marginal benefit will be gained at each step by adding the next most important modality. We will justify this particular ordering of modalities in section 5B, but we defer this discussion in order to present the overall results as concisely as possible. For now, we simply point out that the ordering of modalities is acoustic (most important), ASR, system, visual, and beamforming (least important).

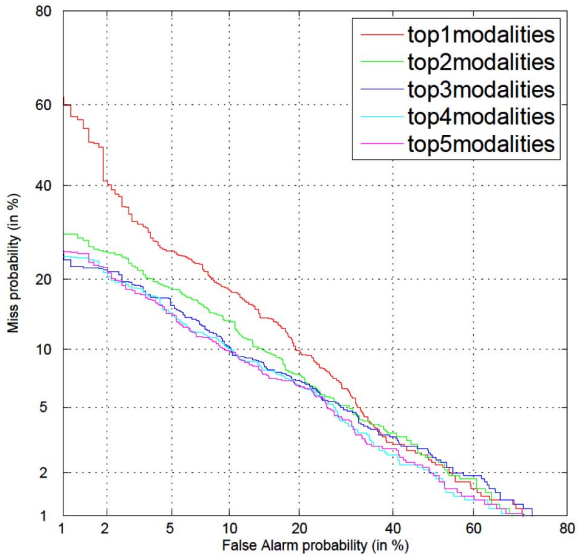

Fig. 3. DET curves showing the incremental improvement by modality for the Adaboost classifier. The order of modalities is acoustic (most important), ASR, system, visual, and beamforming (least important). Each curve shows the performance when features from the top N modalities are used.

There are three observations to make about Fig. 2. First, there is significant improvement by including multiple modalities. For example, the Adaboost classifier improves its EER from 13.9% with one modality to 9.8% with all five modalities. With the random forest classifier, the EER improves from 14.5% to 9.9%. Second, the ensemble classifiers (random forest and Adaboost) show more robustness to overfitting on the modality level. We can clearly see that the regression tree shows overfitting beyond the top three modalities. Just as it is possible to overfit the data with features, we see a similar phenomenon of overfitting with modalities. In this case, using more modalities adds more noise than useful information. In contrast, the ensemble classifiers show consistent but diminishing gains. One of the benefits of ensemble methods is that they tend to be more robust to overfitting. Third, the general ordering of classifiers by performance is Adaboost (best), random forest, logistic regression, and then regression tree (worst). We can see that the regression tree is consistently the worst. Logistic regression and random forest both perform fairly well, but are not as consistent. Adaboost has the most consistent and competitive performance. For this reason, we focus our attention on the Adaboost model in the remainder of our analyses.

B. Performance of Best Classifier

Next we examine more closely the performance of our best-performing classifier: Adaboost. Whereas EER reflects the performance of a system at a single operating point, DET curves characterize the performance across a whole range of operating points. Fig. 3 shows the DET curves for the Adaboost classifier when incrementally adding modalities, again in order of their individual performance. These five DET curves show the full performance characteristic for the rightmost bar in each group in Fig. 2.

We highlight two observations about Fig. 3. First, adding more modalities yields increasingly smaller gains. Only the top three modalities (acoustic, ASR, system) yield statistically significant incremental improvements in EER at a $p < .01$ level of significance. In this case, it may not be worth the effort to compute visual features for marginal gains. We often see the law of diminishing returns when combining features and combining systems, and here we see diminishing returns with combining modalities as well. Second, the performance of the system with the single best modality (acoustic) has very poor performance in the low false-alarm region. Note that all 5 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 2 or more modalities significantly improves system performance.

Overall, we can summarize the benefit of multimodal addressee detection as follows. The best-performing classifier reduces the EER from 13.9% with the single best modality (acoustic features only) to 9.8% with all five feature modalities. Beyond the top three modalities (acoustic, ASR, system), using additional modalities yields little to no benefit.

V. Analysis

In this section we investigate which features and modalities contribute most to overall performance. We will start by assessing the importance of individual features, and then turn our attention to the importance of modalities in aggregate. For these analyses, we again focus on our single best-performing classifier, Adaboost.

A. Importance of Individual Features

For any classifier architecture it is not always clear how to quantify the importance of individual features. One useful metric in the case of Adaboost is the concept of relative influence [43], [39]. Relative influence is the reduction in the loss function attributable to a single feature, normalized by the total reduction in loss due to all features. This measure indicates 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.

Fig. 4 shows the relative influence of the top 30 features in the Adaboost model, sorted in decreasing order. These top 30 features account for more than 95% of the total relative influence. The names of the features have been color coded by modality for ease of interpretation.

Perhaps the most enlightening thing we can do is to simply look at what the 10 most influential features are. These 10 features make up more than 80% of the total relative influence. Here, we explain what the features are, as follows:

1) log likelihood ratio of two energy contour GMMs;
2) log likelihood ratio of two language models;
3) log likelihood ratio of two energy contour GMMs, normalized by the length of audio;
4) confidence of top ASR hypothesis;
5) average energy during the utterance minus the average energy during the 1 sec interval after the utterance;

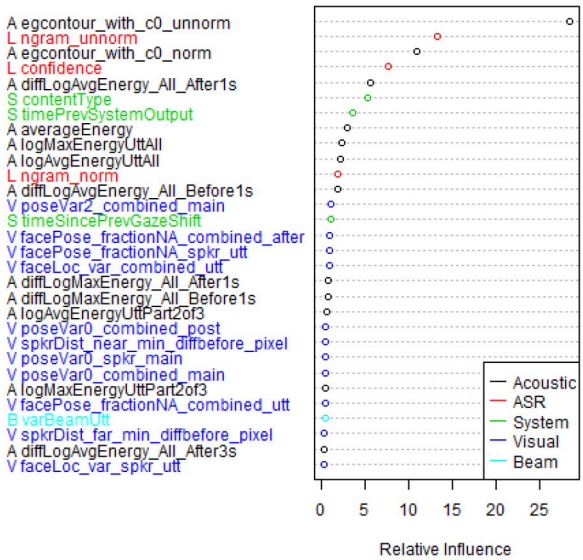

Fig. 4. Relative influence of top 30 features in Adaboost model. The names of the features have been color coded by modality for ease of reference. The modality is also indicated by the first letter of the feature name for those reading in black and white (A = Acoustic, L = ASR).

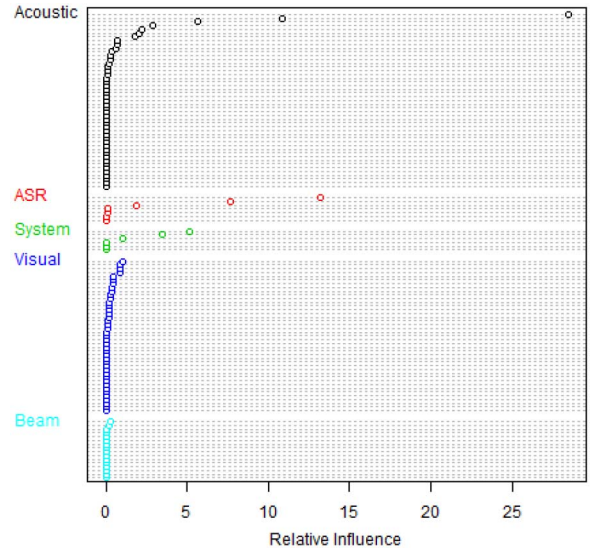

Fig. 5. Relative influence of all 117 features in the Adaboost model. The features are grouped first by modality, then in decreasing order of influence.

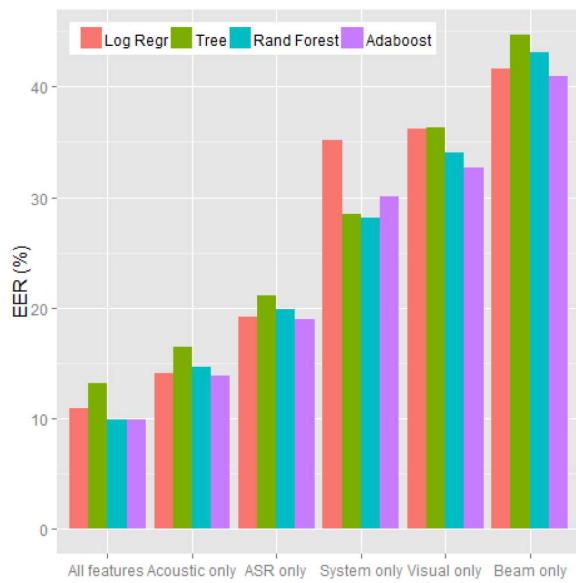

Fig. 6. Equal error rates of systems when only one feature modality is used.

6) dialogue act type of the last computer agent speech (question, confirmation, answer, etc.);
7) time elapsed since the last computer agent speech;
8) average energy during utterance;
9) max frame-level energy during utterance; and
10) log average energy during utterance.

Looking at the list of top 10 features above, we can make a few observations. First, acoustic features dominate. Six of the top 10 features are related to acoustic energy. Second, ASR contributes in the form of n-gram likelihood ratios and confidence score. Third, context helps. Some of the top 10 features have to do with what or when the computer agent last spoke, or the energy level immediately after the utterance. These features capture information outside of the time interval in which the utterance was actually spoken. And finally, no beamforming or visual features appear in the top 10 feature list. This suggests that these modalities are much less useful for this task, confirming the results from Section IV. It may seem surprising that visual features such as face-pose angle are not very important, and we will explain the main reasons for this in the discussion section.

Finally we note that the three most important features are the model-based prosodic and lexical features that were found to be highly effective in recent work by Shriberg *et al.* [5], [7]. The fact that these features perform similarly here, on an entirely different data set, motivates examining the marginal benefit of other modalities when added to the acoustic and ASR-based features.[4]

B. Importance of Different Modalities

In addition to knowing the relative importance of individual features, we would also like to know the relative importance of different modalities. We will approach this using three different methods.

The first approach is to visualize the importance of individual features when grouped by modality. Fig. 5 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. The top few ASR and system features are useful. The rest don't seem to contribute much. Note that the ordering of modalities suggested by Fig. 5 matches the ordering in Fig. 3.

The second approach is to run full end-to-end experiments using one feature modality at a time. These experiments reveal how well we can do addressee detection when only using information from a single modality. Fig. 6 shows a barplot of EERs

[4]The results obtained here with energy contour features alone (about 19% EER, cf. Table II), are remarkably close to the 17% EER obtained on the corpus used in [7].

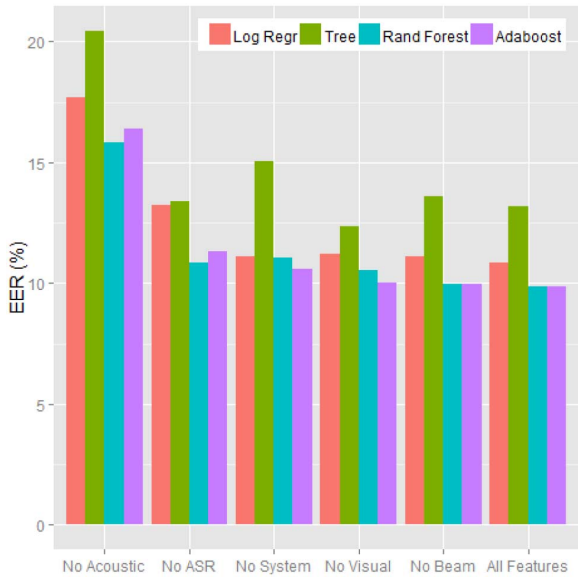

Fig. 7. Equal error rates of systems when one feature modality is removed.

for our leave-one-group-in experiments. The leftmost group of bars shows the performance when all feature modalities are used, as a reference.

There are several things to notice in Fig. 6. The group EER increases as we go from left to right. This trend holds true regardless of the classifier. This suggests that the order of importance among the various modalities is acoustic (most important), ASR, system, visual, and beamforming (least important). Also, using multiple modalities helps substantially. The leftmost group of bars is much lower than any other group of bars. No matter which single modality you pick, there is significant benefit to using multiple modalities. Again, note that the ordering of modalities suggested by Fig. 6 matches the ordering in Fig. 3.

The third approach is to run full end-to-end experiments omitting one group of features at a time. These experiments will reveal how much removing a particular modality from the full multimodal system negatively affects system performance. Fig. 7 shows a barplot of EERs for our leave-one-group-out experiments. The rightmost group of bars shows the performance of the full multimodal feature set, as a reference. Since we are measuring the effect of *removing* a modality, a higher EER indicates that the modality is more important. Higher bars mean greater importance.

Fig. 7 generally supports our other findings. The results get much worse when we remove acoustic features, suggesting that they are very important. The results are somewhat worse when we remove ASR or system features, suggesting that they are moderately important. The results are not negatively affected much when we remove the visual or beamforming features. In fact, the results actually get better in some cases due to overfitting with the regression tree. This suggests that these feature modalities are not very important. For our Adaboost model, only leaving out energy or ASR features yielded a statistically significant increase in EER at a $p < .01$ level of significance, and leaving out system features yielded a statistically significant increase at a $p < .05$ level of significance. Importantly,

TABLE II
EQUAL ERROR RATES (IN%) OF ADABOOST MODEL WHEN WE STRIP AWAY ABSOLUTE ENERGY INFORMATION. THE TWO MIDDLE COLUMNS SHOW THE RESULTS WHEN WE LEAVE ONE FEATURE MODALITY IN OR TAKE ONE FEATURE MODALITY OUT. THE RIGHTMOST COLUMN SHOWS THE RESULTS OF INCREMENTALLY ADDING FEATURE MODALITIES FROM TOP TO BOTTOM

Features	Leave-in	Leave-out	Incremental
ASR	19.01	16.39	19.01
Acoustic	19.16	16.39	15.27
System	30.02	14.50	15.37
Visual	32.68	14.65	14.60
Beamforming	40.98	14.60	**13.99**

leave-one-in and leave-one-out experiments arrive at the same ordering of modalities, which justifies the ordering of importance used in Section IV.

Putting all our analysis experiments together, we can characterize the relative importance of the individual modalities as follows: Acoustic features are extremely important. ASR and system features have medium importance. Visual and beamforming features have little to no importance.

C. Additional Analyses

We have seen that acoustic energy-based features are very important in predicting addressee for the given experimental setup. With a view toward generality and future applications, we may not want our addressee detection system to depend on the users speaking more loudly when addressing the system. This dependence on differing vocal effort is an artifact of the computer agent's limited capabilities, and may be a barrier to more natural interaction with the agent. We therefore tried to determine how our system would perform if we removed dependence on utterance-level energy. This analysis anticipates the ongoing development of dialogue systems which allow humans to speak to the system in a more natural manner. To this end, we repeated several of the above experiments and analyses, but excluding features that depend on absolute energy levels. We also modified the energy contour models to omit the first DCT value describing c0 energy. This effectively removes all of the acoustic features except those that model speaking style as expressed by utterance-internal energy variation [5].

Table II shows the EERs of the Adaboost model without absolute energy information. The second column shows the system performance when only a single feature modality is used. The third column shows the system performance when a single feature modality is omitted. The rightmost column shows the results of incrementally adding feature modalities from top to bottom. So, the top entry refers to using ASR features only, and the bottom entry refers to using all five feature modalities. The result in the bottom row, right-most column has the performance with all five modalities, and therefore serves as a reference point. The numbers in Table II show a significant drop in performance compared to the system with absolute energy information. For example, with only acoustic features, the EER increases from 13.9% ("top1modalities" curve in Fig. 3) to 19.2% (acoustic only in Table II), a relative change of 38%. Similarly, with all modalities, the EER increases from 9.8% ("top5modalities" curve in Fig. 3) to 14.0% (bottom right entry of Table II), a relative change of 43%.

Several additional observations should be noted. First, the leave-one-group-in experiments suggest the same ordering of modalities as before, except that acoustic features have fallen from being most important to being second in importance behind ASR features. Second, ASR and acoustic features as a group (both based on audio input) remain dominant. The leave-one-in and leave-one-out experiments indicate that these two modalities are dominant in importance compared to the other three modalities. Only these two modalities yield statistically significant changes to EER at a $p < .01$ level of significance in the leave-one-out experiments. Similarly, only these two modalities yield statistically significant incremental improvement to EER at a $p < .01$ level of significance in the incremental-by-modality experiments. Third, the visual and beamforming modalities seem to contribute more when absolute energy information is stripped away. Whereas before these two modalities did not benefit system performance almost at all, here we see that they play a more important role when either omitted or added incrementally.

Another question of interest is to determine how useful it is to consider the context *after* an utterance is spoken. As described earlier, several of our features look at intervals of time up to 3 seconds after an utterance is spoken. Are these features actually useful? When we look at the Adaboost model with the full set of 117 multimodal features, we find that the 30 features which use post-utterance information constitute about 9.4% of the total relative influence. Of these 30 features, 7 consider time intervals up to 1 second after the utterance ends, and these 7 features constitute 6.4% of the total relative influence, more than $\frac{2}{3}$ of the post-utterance combined influence. So, it seems that there is some useful information contained in the interval after an utterance is spoken, but most of this information is concentrated in the interval immediately after the utterance ends. We can also verify this hypothesis by observing the effect of removing post-utterance features. When we remove all 30 post-utterance features, the EER of the Adaboost classifier increases from 9.8% to 10.5%. However, when we remove only the 23 features not contained in the "utterance + 1 sec" subset, the EER only increases from 9.8% to 9.9%. This result is useful, since it suggests that significant performance benefits may be gained from limited post-utterance information, while still operating within a real-time, low-latency scenario. A half-second or one-second delay may be acceptable in many real-time applications, whereas a 3-second delay would almost certainly be unacceptable.

D. Discussion

The fact that visual features were not very important may seem surprising. Because it is such an important cue in H-H interactions, one would naturally assume that it would be an important cue in H-H-C interactions as well. The main reason for this result is that the computer was a major situational attractor: people continued looking at the screen even when they were talking amongst themselves, a phenomenon that has been observed and studied in social psychology [44]. For the given scenario, we can characterize the face orientation as "necessary but not sufficient." In other words, if a person is not looking at the computer, they are almost certainly not addressing the system.

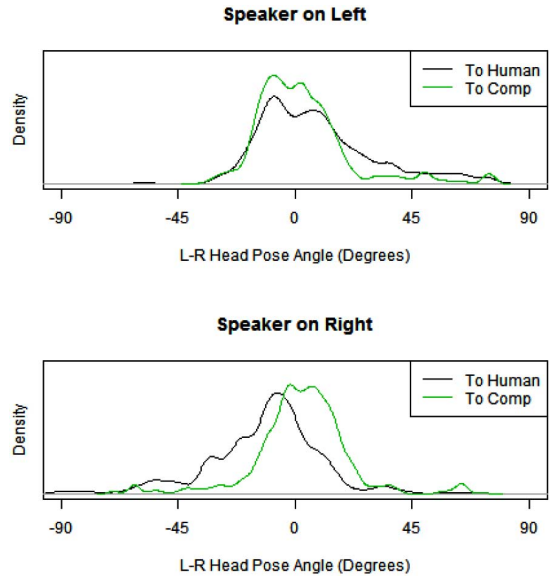

Fig. 8. Distribution of left-right head pose angle of speaker, separated by speaker location and intended addressee. This only includes data for which there is a single speaker and single addressee, and the speaker is located on the left or the right (not in the center).

But if a person *is* looking at the computer, they may or may not be addressing the system. Because people's default face orientation for this task was towards the computer, face pose provides little useful information. The strength of a situational attractor may vary widely depending on the context, but the general phenomenon has been observed in other studies [17], [18]. Another contributing factor is that the face-pose estimates were not robust enough to be useful. The pose estimates were fairly noisy and could only track face angles within a limited range. We can take measures to partially compensate for these shortcomings (such as measuring the fraction of time that pose estimates are unavailable), but they did not seem to provide enough reliable information to be useful. In some instances, however, head pose angle might be useful. Fig. 8 shows the distribution of the speaker's head pose angle in the left-right direction for the subset of training data satisfying all of the following conditions: (1) there is a single speaker, (2) there is a single intended addressee, and (3) the speaker is located on the left or right side (not in the center). We can see that there is some separation in the distributions that might be exploited if we partition the data according to appropriate criteria.

Some of the most important features are features that we may not want to rely upon. Acoustic energy was a very important source of information in our data set, but it relies on the fact that people speak differently to a computer than to another person. As conversational systems become more and more natural, we may not want to rely on people speaking in a distinctly different manner towards computers. Similarly, ASR confidence was a very useful feature, but is highly sensitive to the acoustic environment, language model coverage, nonnative accents, and other incidental factors affecting the recognition system.

Features describing context should be explored in more depth. We saw that what the computer agent last said and when the computer agent last spoke are useful indicators. We also saw that features describing the time intervals immediately

after the utterance were informative. These features all describe the *context* in which the utterance is spoken. Dialogue context is also readily and reliably available to the system, especially since much of it is produced by the system itself (such as the dialogue act or display contents last generated). Context features are therefore attractive to use for addressee detection, and merit further exploration. One potential drawback is that they are highly specific to the system task, i.e., it would hard to incorporate them in a general way in a "black box" addressee detector, at least without retraining and calibration of the overall system. It might be possible to develop a descriptive schema for dialogue context that generalizes across systems and allows sharing of training data, which would facilitate the development of new H-H-C systems.

VI. Conclusion

We have proposed a multimodal addressee-detection system that uses acoustic, visual, system state, beamforming, and ASR information. Using data from a multiparty dialogue scenario, we assess its performance and determine which types of information are most useful in predicting addressee. We find that acoustic information is most useful, dominating the other modalities in importance. Lexical and dialogue state information are also useful, providing significant performance gains. Visual and beamforming information provide little to no additional benefit. Our findings are relevant to multiparty, open-world dialogue systems in which a computer agent plays an active role in structuring the conversation. For these situations, our experiments suggest that audio-based 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. Our analyses also suggest that as dialogue systems become more and more natural, the acoustic information will become less dominant, though still important, while the other modalities will increase in relative importance.

There are two primary avenues for future work. The first is to explore a richer set of features that describe the *context* in which a person speaks–what the user is responding to and how their response fits into the larger, overarching interaction. The second avenue is to collect a rich multimodal data set in a human-human-computer scenario where the user is the active initiator, rather than the computer. Such a data set with addressee annotations would enable further study on how to design effective and general automatic addressee detection systems.

Acknowledgment

The authors would like to thank D. Bohus and O. Vinyals for help with the data set and feedback on the paper, and L. Ferrer for advice on statistical tests.

References

[1] J. Dowding, R. Alena, W. J. Clancey, M. Sierhuis, and J. Graham, "Are you talking to me? Dialogue systems supporting mixed teams of humans and robots," in *Proc. AAAI Fall Symp.: Aurally Informed Performance: Integrating Mach. Listening Auditory Presentation Robot. Syst.*, Washington, DC, USA, Oct. 2006, pp. 22–27.

[2] T. Paek, E. Horvitz, and E. Ringger, "Continuous listening for unconstrained spoken dialog," in *Proc. ICSLP*, B. Yuan, T. Huang, and X. Tang, Eds., Oct. 2000, vol. 1, pp. 138–141.

[3] D. Bohus and E. Horvitz, "Multiparty turn taking in situated dialog: Study, lessons, and directions," in *Proc. ACL SIGDIAL*, Jun. 2011, pp. 98–109.

[4] O. Vinyals, D. Bohus, and R. Caruana, "Learning speaker, addressee and overlap detection models from multimodal streams," in *Proc. 14th ACM Int. Conf. Multimodal Interaction*, Oct. 2012, pp. 417–424.

[5] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Proc. Interspeech*, Sep. 2012, pp. 334–337.

[6] H. Lee, A. Stolcke, and E. Shriberg, "Using out-of-domain data for lexical addressee detection in human-human-computer dialog," in *Proc. North Amer. ACL/Human Language Technol. Conf.*, Jun. 2013, pp. 221–229.

[7] E. Shriberg, A. Stolcke, and S. Ravuri, "Addressee detection for dialog systems using temporal and spectral dimensions of speaking style," in *Proc. Interspeech*, Aug. 2013, pp. 2559–2563.

[8] B. Brumitt and J. J. Cadiz, "Let there be light: Examining interfaces for homes of the future," in *Proc. Int. Conf. Human Comput. Interaction*, Aug. 2001, pp. 375–382.

[9] P. P. Maglio, T. Matlock, C. S. Campbell, S. Zhai, and B. A. Smith, "Gaze and speech in attentive user interfaces," in *Advances in Multimodal Interfaces–ICMI 2000*, T. Tan, Y. Shi, and W. Gao, Eds. New York, NY, USA: Springer, Oct. 2000, pp. 1–7.

[10] M. K. Lee, S. Kiesler, and J. Forlizzi, "Receptionist or information kiosk: How do people talk with a robot?," in *Proc. 2010 ACM Conf. Comput. Supported Cooperative Work*, 2010, pp. 31–40.

[11] A. Knott and P. Vlugter, "Multi-agent human–machine dialogue: Issues in dialogue management and referring expression semantics," *Artificial Intelligence*, vol. 172, no. 2, pp. 69–102, 2008.

[12] R. Lunsford and S. Oviatt, "Human perception of intended addressee during computer-assisted meetings," in *Proc. 8th Int. Conf. Multimodal Interfaces*, 2006, pp. 20–27.

[13] J. Terken, I. Joris, and L. De Valk, "Multimodal cues for addressee-hood in triadic communication with a human information retrieval agent," in *Proc. 9th Int. Conf. Multimodal Interfaces*, 2007, pp. 94–101.

[14] I. Bakx, K. Van Turnhout, and J. Terken, "Facial orientation during multi-party interaction with information kiosks," in *Proc. INTERACT*, Sep. 2003, pp. 163–170.

[15] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 144–151.

[16] N. Baba, H.-H. Huang, and Y. I. Nakano, "Addressee identification for human-human-agent multiparty conversations in different proxemics," in *Proc. 4th Workshop Eye Gaze Intell. Human Mach. Interaction*, 2012, pp. 6:1–6:6.

[17] H.-H. Huang, N. Baba, and Y. Nakano, "Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation using nonverbal information," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 401–408.

[18] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen, "Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features," in *Proc. 7th Int. Conf. Multimodal Interfaces*, 2005, pp. 175–182.

[19] G. Skantze and J. Gustafson, "Attention and interaction control in a human-human-computer dialogue setting," in *Proc. SIGDIAL 2009 Conf.: 10th Annu. Meeting Special Interest Group Discourse Dialogue*, Stroudsburg, 2009, pp. 310–313.

[20] T. J. Tsai, A. Stolcke, and M. Slaney, "Multimodel addressee detection in multiparty dialogue systems," in *Proc. ICASSP*, Apr. 2015, pp. 2314–2318.

[21] D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *Proc. Int. Conf. Multimodal Interfaces Workshop Mach. Learn. Multimodal Interaction*, 2010, pp. 5:1–5:8.

[22] C. Zhang and Y. Rui, "Robust visual tracking via pixel classification and integration," in *Proc. Int. Conf. Pattern Recog.*, Aug. 2006, vol. 3, pp. 37–42.

[23] Q. Wang, W. Zhang, X. Tang, and H.-Y. Shum, "Real-time bayesian 3-D pose tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 12, pp. 1533–1541, 2006.

[24] D. Bohus and E. Horvitz, "Dialog in the open world: Platform and applications," in *Proc. Int. Conf. Multimodal Interfaces*, 2009, pp. 31–38.

[25] D. DeVault *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2014, pp. 1061–1068.
[26] D. Bohus, C. W. Saw, and E. Horvitz, "Directions robot: In-the-wild experiences and lessons learned," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2014, pp. 637–644.
[27] "Tobii REX technical specifications and FAQ," Tobii Technology AB. Danderyd , Sweden, 2014 [Online]. Available: http://developer.tobii.com/rex-technical-specs-faq/
[28] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[29] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.
[30] D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, and J.-F. Ladry, "Fusion engines for multimodal input: A survey," in *Proc. Int. Conf. Multimodal Interfaces*, 2009, pp. 153–160.
[31] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, Jun. 2011, pp. 689–696.
[32] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inform. Process. Syst.*, 2012, pp. 2222–2230.
[33] H.-I. Suk, S.-W. Lee, D. Shen, and A. D. N. Initiative, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.
[34] H. P. Martínez and G. N. Yannakakis, "Deep multimodal fusion: Combining discrete events and continuous signals," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 34–41.
[35] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, "Multimodal fusion using dynamic hybrid models," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 556–563.
[36] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. Assoc. Comput. Linguistics*, Aug. 2013, pp. 973–982.
[37] D. Freedman, *Statistical Models: Theory and Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[38] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Pacific Grove, CA, USA: Wadsworth and Brooks, 1984.
[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[40] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J. Jpn. Soc. Artificial Intell.*, vol. 14, no. 1612, pp. 771–780, 1999.
[41] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and Przybocki, "The DET curve in assessment of detection task performance," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. ADA530509, 1997 [Online]. Available: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA530509
[42] L. Ferrer, "Statistical modeling of heterogeneous features for speech processing tasks," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, Dec. 2008 [Online]. Available: http://www.speech.sri.com/people/lferrer/thesis.html
[43] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, pp. 1189–1232, 2001.
[44] M. Argyle and J. A. Graham, "The central europe experiment: Looking at persons and looking at objects," *Environ. Psychol. Nonverbal Behavior*, vol. 1, no. 1, pp. 6–16, 1976.

T. J. Tsai (S'13) received the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2006 and 2007, respectively, and is currently working toward the Ph.D. degree in electrical engineering and computer science at the University of California at Berkeley, Berkeley, CA, USA.

From 2008 to 2010, he was with SoundHound, Santa Clara, CA, USA.

Andreas Stolcke (M'96–SM'05–F'11) received the Ph.D. degree in computer science from the University of California at Berkeley, Berkeley, CA, USA.

He was previously a Senior Research Engineer with the Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA. He is currently a Principal Researcher with the Speech and Dialog Research Group, Microsoft Research, Mountain View, CA, USA. He is also an External Fellow with the International Computer Science Institute, Berkeley, CA, USA. He authored a widely used open-source toolkit for statistical language modeling. His research interests include machine language learning, parsing, speech recognition, speaker recognition, and speech understanding.

Malcolm Slaney (S'78–M'84–SM'01–F'10) is a Research Scientist with the Machine Hearing Group, Google Research, Mountain View, CA, USA. He is also a Consulting Professor with the Center for Computer Research in Music and Acoustics, Department of Music, Stanford, CA, USA, and an Affiliate Faculty Member with the Electrical Engineering Department, University of Washington, Seattle, WA, USA. Before joining Google, he was with: AT&T's Bell Laboratories, Naperville, IL, USA, and Holmdel, NJ, USA; Schlumberger Palo Alto Research, Palo Alto, CA, USA; Apple Computer, Cupertino, CA, USA; Interval Research, Palo Alto, CA, USA; IBM's Almaden Research Center, San Jose, CA, USA; Yahoo! Research, Sunnyvale, CA, USA; and Microsoft Research, Mountain View, CA, USA. He is a coauthor of the book *Principles of Computerized Tomographic Imaging* (IEEE Press, 2008), which was republished by SIAM in their Classics in Applied Mathematics Series. He is coeditor of the book *Computational Models of Auditory Function* (IOS Press, 2001). His current research interests include understanding the role of attention in conversational speech and general audio perception.

Dr. Slaney was an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND SIGNAL PROCESSING and the *IEEE MultiMedia Magazine*. He has given successful tutorials on applications of psychoacoustics to signal processing at ICASSP 1996 and 2009, on multimedia information retrieval at SIGIR and ICASSP, and on WebScale Multimedia Data at ACM Multimedia 2010.