Latent Mixture of Discriminative Experts

Derya Ozkan and Louis-Philippe Morency
University of Southern California
Institute for Creative Technologies
{ozkan,morency}@ict.usc.edu

***Abstract*—In this paper, we introduce a new model called Latent Mixture of Discriminative Experts which can automatically learn the temporal relationship between different modalities. Since, we train separate experts for each modality, LMDE is capable of improving the prediction performance even with limited amount of data. For model interpretation, we present a sparse feature ranking algorithm that exploits L_1 regularization. An empirical evaluation is provided on the task of listener backchannel prediction (i.e head nod). We introduce a new error evaluation metric called User-adaptive Prediction Accuracy that takes into account the difference in people's backchannel responses. Our results confirm the importance of combining five types of multimodal features: lexical, syntactic structure, part-of-speech, visual and prosody. Latent Mixture of Discriminative Experts model outperforms previous approaches.**

***Index Terms*—Multimodal integration, mixture of experts, backchannel feedback, multimodal prediction models, evaluation metric, sparse regularization**

I. Introduction

Along with the advances in multimodal systems and interfaces (i.e. smartphones, Microsoft Kinect), processing of multimodal information has gained great attention by many researchers. One of the main problems of multimodal information processing includes effective and efficient fusion of modalities from multiple resources. If integrated carefully, different modalities can provide complementary information that improves the performance of a system.

While earlier work focused on either feature or decision fusion, new models have emerged that are specifically designed for multimodal data. There are several characteristics that a good fusion process is desired to have. Among others, we discuss three of the most important characteristics. First, a good fusion process should be able to allow reweighting of noisy channels. In other words, it should be able to learn how confident each modality is in achieving a defined task (such as audio-visual speaker detection, human tracking, etc.). Second, effective training should be possible, even with limited amount of data. And third, fusion process should be interpretable, therefore analysis of each media should be made feasible.

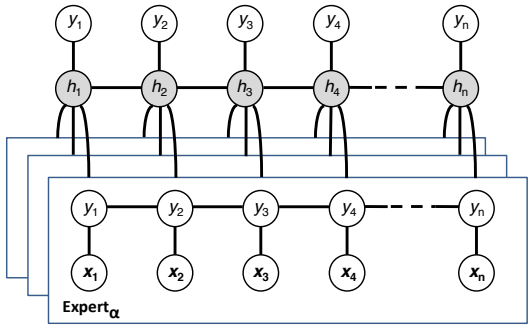


Fig. 1. Latent Mixture of Discriminative Experts: a new dynamic model for multimodal fusion. In this graphical representation, x_j represents the j^{th} multimodal observation, h_j is a hidden state assigned to x_j , and y_j the class label of x_j. Gray circles are latent variables. The micro dynamics and multimodal temporal relationships are automatically learned by the hidden states h_j during the learning phase.

In this paper we introduce a new model called Latent Mixture of Discriminative Experts (LMDE), which directly addresses these three issues. A graphical representation of LMDE is given in Figure 1. One of the main advantages of our computational model is that it can automatically discover the hidden structure among modalities and learn the dynamic between them. Since a separate expert is learned for each modality, effective training can be purveyed even with limited amount of data. Furthermore, our learning process provides a ground for better model interpretability. By analyzing each expert, the most important features in each modality– relevant to the task– can be conceived. To enable

efficient feature analysis, we propose a sparse feature ranking scheme based on L_1 regularization technique [1], [2], [3].

We present empirical evaluation on the task of backchannel feedback prediction confirming the importance of combining different types of multimodal features. Backchannel feedbacks include nods and para-verbals such as "uh-huh" and "mm-hmm" that listeners produce as they are speaking. Predicting when to give backchannel feedback is a good example of complementary information, for which people naturally integrate speech, gestures and higher level linguistic features. Figure 2 shows an example of backchannel prediction where a listener head nod is more likely. These prediction models have broad applicability, including the improvement of nonverbal behavior recognition, the synthesis of natural animations for robots and virtual humans, the training of cultural-specific nonverbal behaviors, and the diagnoses of social disorders (e.g., autism spectrum disorder).

One last issue directly addressed in this paper is the evaluation metric for our multimodal prediction model. Listener feedback varies among people and is often optional (listeners can always decide to give feedback or not). Therefore, traditional error measurements (i.e. recall, precision, f-score) may not always be adequate to evaluate the performance of a prediction model. In this paper, we propose a new error measurement called User-adaptive Prediction Accuracy (UPA) which takes into account the differences in people's nonverbal responses.

Our experiments are performed on a dataset of 45 storytelling dyadic interactions [4][1]. We compare our LMDE model with previous approaches based on Conditional Random Fields (CRF) [5], Latent-Dynamic CRFs [6], and CRF Mixture of Experts (a.k.a Logarithmic Opinion Pools [7]), and a rule based random predictor [8]. All the results are validated by our User-adaptive Prediction Accuracy as well as the traditional error measurements like F1-score. We also provide an analysis of the most important features for each modality and give an intuition on why our intermediate fusion approach improves prediction performance.

The rest of the paper is organized as follows. We first present the related works in Section II. Then we present our Latent Mixture of Discriminative Experts model in Section III, and sparse feature ranking scheme in Section IV. We discuss the challenges in multimodal prediction modeling and describe our error computation metric in Section V. Experimental setup is explained in Section VI, Results and discussions are given in Section VII. Finally, we conclude with future research directions in Section VIII.

[1]Freely available at http://rapport.ict.usc.edu/

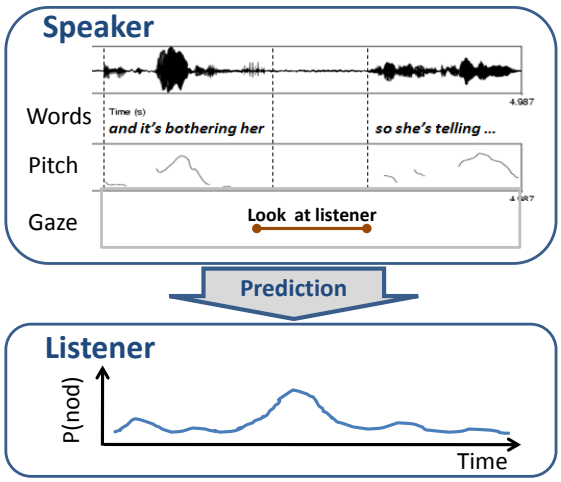


Fig. 2. Example of multimodal prediction model: listener nonverbal backchannel prediction based on speaker's speech and visual gestures. As the speaker says the word *her*, which is the end of the clause (*her* is also the object of the verb *bothering*), and lowers the pitch while looking back at the listener and eventually pausing, the listener is then very likely to head nod (i.e., nonverbal backchannel).

II. Related Work

Multimodal information processing can be achieved mainly in three levels: early, late and intermediate [9]. Early fusion involves feature level integration, which exploits the correlation among all features [10], [11], [12]. McCowan et al. [13] presented a multimodal approach for recognition of group actions in meetings. In their experiments, early integration gives significantly better frame error rates than all approaches apart from audio-visual Asynchronous Hidden Markov Model system, which is used to model the interactions between individuals. However, modeling temporal synchrony/asynchrony among modalities is a hard problem in early fusion, since features from different modalities do not always happen at the same time.

On the other hand, late fusion refers to decision level integration, in which the decisions of individual modalities are fused together to have a final

decision [14], [15], [16]. This level of integration is usually more scalable than feature level integration, since the decisions from multiple media are all in the same format. Snoek et al. [17] compares early fusion and late fusion for semantic concept learning from multimodal video. In their experiments, late fusion gives better performance for most concepts; however it comes with a cost of increased learning effort. For both early and late fusion, classifiers are generic, which are also used for unimodal data processing.

In this paper, we present a probabilistic model (i.e. intermediate fusion) specifically designed for multimodal fusion, where the integration is done at the model level. Factorial Hidden Markov Models [18], Coupled Hidden Markov Models [18] and Layered Hidden Markov Models (LHMMs) are examples of statistical models for intermediate fusion of audio visual data. LHMMs was proposed in [19] for modeling office activity from multiple sensory channels. LHMMs can be seen as a cascade of Hidden Markov Models, where each layer is trained independently, and the results from a lower layer are used as input to an upper layer. Barnard and Odobez [20] use this framework in combination with unsupervised clustering of the data for event recognition in sports videos. Different than earlier intermediate fusion techniques, our model depends on discriminative models that can learn the dynamic among different modalities.

Jordan et. al. [21] presented the Hierarchical Mixture of Experts (HME) based on probabilistic splits of the input space. HME models a mixture of component distributions referred to as experts, where the expert mixing ratios are set by gating functions. Bishop et. al. [22] proposed a variant of this model called Bayesian HME (BME) based on variational inference. HME and BME are mainly used for solving static regression and classification problems. On the other hand, we propose a discriminative model for solving sequential patterns, where we predict one label per time sample. Sminchisescu et. al. [23] used the BME approach for discriminative inference in continuous chain models. Similar to our LMDE model, it can learn the mixing coefficients among experts. In addition to this, LMDE exploits a latent variable that allows multiple mixing coefficients. In other words, each hidden state in our LMDE model can represent a different set of mixing coefficients.

The application described in this paper integrates multimodal cues from one person are used to predict the social behavior of another participant. This type of predictive models has been mostly studied in the context of embodied conversational agents [24], [25]. Several researchers have developed models to predict when backchannel should happen. In general, these results are difficult to compare as they utilize different corpora and present varying evaluation metrics. Ward and Tsukahara [8] propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110ms during speech. Models were produced manually through an analysis of English and Japanese conversational data. Fujie et al. [26] use Hidden Markov Models to perform head nod recognition. In their paper, they combined head gesture detection with prosodic low-level features from the same person to determine strongly positive, weak positive and negative responses to yes/no type utterances.

Maatman et al. [27] present a multimodal approach where Ward and Tsukhara's prosodic algorithm is combined with a simple method of mimicking head nods. No formal evaluation of the predictive accuracy of the approach was provided but subsequent evaluations have demonstrated that generated behaviors do improve subjective feelings of rapport [28] and speech fluency [4]. Morency et al. [29] showed that Conditional Random Field models can be used to learn predictive features of backchannel feedback. In their approach, multimodal features are simply concatenated in one large feature vector for the CRF model. They show statistical improvement when compared to the rule-based approach of Ward and Tsukahara [8]. Our experiments described in Section VI compare with this early fusion approach.

Feature selection refers to the task of finding a subset of features that are most relevant to the model, and provides a good representation of data. It alleviates the problem of overfitting by eliminating the noisy features. With only the relevant features, a better understanding and analysis of data is facilitated. A well known feature selection technique based on L_1-regularization was applied for conditional random fields in robot tag domain [3]. Based on the gradient-based feature selection method (grafting) in [30], Vail et. al. [31] proposed an incremental feature selection technique for Maximum Entropy Modeling. For the task of listener backchannel prediction, Morency et. [29]

proposed a greedy approach where the first feature is selected based on it's performance on the task when used individually. Then, new features are selected incrementally based on their effect in the performance when added to the first feature. Different than this greedy approach, all features are present during the selection process in our sparse feature ranking scheme.

The three main contributions of this paper are:

- Latent Mixture of Discriminative Experts model for multimodal data integration.
- A sparse feature ranking scheme for expert data analysis.
- User-adaptive Prediction Accuracy for better evaluation.

III. Latent Mixture of Discriminative Experts

The task of multimodal prediction involves effective and efficient fusion of information from multiple sources. One of the desired characteristics of good prediction model is that it should be able learn the temporal relationships between modalities. In this paper, we introduce a multimodal fusion algorithm called Latent Mixture of Discriminative Experts (shown in Figure 1), that addresses important challenges involved in multimodal data processing. (1) The hidden states of LMDE can automatically learn the hidden dynamic between modalities. (2) By training separate experts, we improve the prediction performance even with limited amount of data. (3) LMDE provides interpretability of modalities, which can be accomplished by expert analysis.

The task of our LMDE model is to learn a mapping between a sequence of multimodal observations $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ and a sequence of labels $\mathbf{y} = \{y_1, y_2, ..., y_m\}$. Each y_j is a class label for the j^{th} frame of a video sequence and is a member of a set $\mathcal{Y}$ of possible class labels, for example, $\mathcal{Y} = \{\texttt{backchannel}, \texttt{no feedback}\}$. Each frame observation x_j is represented by a feature vector $\in \mathbf{R}^d$, for example, the prosodic features at each sample. For each sequence, we also assume a vector of "sub-structure" variables $\mathbf{h} = \{h_1, h_2, ..., h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model. Each h_j is a member of a set $\mathcal{H}_{y_j}$ of possible hidden states for the class label y_j. $\mathcal{H}$, the set of all possible hidden states, is defined to be the union of all $\mathcal{H}_y$ sets.

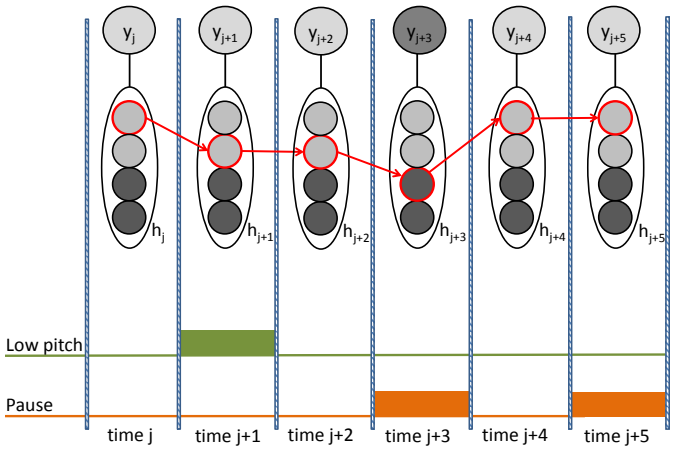


Fig. 3. An example of how the hidden variables of our LMDE model can learn the temporal dynamics and asynchrony between modalities.

In the rest of this section, we first provide some intuitions motivating our model; then present details of our LMDE model, explain how we learn the model parameters and finally how inference is performed.

A. Motivation

To illustrate how our LMDE can use its latent variables to learn the hidden temporal relationship between modalities, we present an example (shown in Figure 3) based on the application of predicting listener responses known as backchannel feedback. In this scenario, the goal is to predict when a listener is most likely to predict a head nod (i.e the label y_{j+3}) given the input features extracted from the speaker actions. In our LMDE model, each source of information (e.g. visual, lexical, auditory) is modeled by an expert. In our example we have two experts: pause/talking (orange) and low pitch region in speech (green). Figure 3 shows the speaker talking with low pitch at time $j+1$. We know from literature [8] that listeners are more likely to give a backchannel feedback (1) during a pause and (2) shortly after a region of low pitch (usually around 700ms after the low pitch region). Our LMDE model can easily learn this temporal asynchronous relationship between speaker pause, speaker low pitch region and listener response by using only two hidden states per label (i.e. $|\mathcal{H}| = 2$).

In Figure 3, the first two hidden states (light gray circles) of each hidden variable h_j are associated with the label *no feedback* and the last two hidden states (dark gray circles) are associated with the

label *backchannel*. At time j, the speaker is talking and none of the experts are active. Then, we see low pitch region at time $j+1$, which activates the hidden state 2. At time $j+2$, the speaker is still talking but with no low pitch region. Remark that since the second hidden state was activated at time $j+1$ by the low pitch region, the same hidden state will stay active[2]. This is an example where the LMDE model shows memory functionality through its hidden variables h_j[3]. At time $j+3$, the hidden state 3 is activated due to a pause in speaker's talk, which triggers prediction of a listener backchannel at that point in time. Then, at time $j+4$ the LMDE model gets back to the hidden state 1 when the speaker starts talking again. No head nod will be predicted at time $j+5$, even though the speaker paused (because no low pitch region occurred earlier). Another important aspect of the LMDE model illustrated in Figure 9 of our experimental results (see Section VII) is that the latent variables h_j can learn multiple mixtures of experts, with one set of mixture weights per hidden state. More details on the LMDE model and the latent variables are given in the following subsections.

B. LMDE Model

Following Morency et al. [6], we define our LMDE model as follows:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}} P(\mathbf{y} \mid \mathbf{h}, \mathbf{x}, \theta) P(\mathbf{h} \mid \mathbf{x}, \theta) \quad (1)$$

where θ are model parameters learned during training.

To keep training and inference tractable, Morency et al. [6] restrict the model to have disjoint sets of hidden states $\mathcal{H}_{y_j}$ associated with each class label. Since sequences which have any $h_j \notin \mathcal{H}_{y_j}$ will by definition have $P(\mathbf{y} \mid \mathbf{h}, \mathbf{x}, \theta) = 0$, the latent conditional model becomes:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}:\forall h_j \in \mathcal{H}_{y_j}} P(\mathbf{h} \mid \mathbf{x}, \theta). \quad (2)$$

where

$$P(\mathbf{h} \mid \mathbf{x}, \theta) = \frac{\exp\left(\begin{array}{c} \sum_l \theta_l \cdot \mathbf{T}_l(\mathbf{h}) + \\ \sum_s \theta_s \cdot \mathbf{S}_s(\mathbf{h}, \mathbf{x}) \end{array} \right)}{\mathcal{Z}(\mathbf{x'}, \theta)}, \quad (3)$$

For convenience, we split θ into two parts: θ_l parameters related to the transition between hidden states, and θ_s parameters related to the relationships between expert outputs and the hidden states h_j. $\mathcal{Z}$ is the partition function, and $\mathbf{T}_l(\mathbf{h}, \mathbf{x'})$ is defined as follows:

$$\mathbf{T}_l(\mathbf{h}) = \sum_j t_l(h_{j-1}, h_j, j), \quad (4)$$

where j corresponds to the frame index, and $t_l(h_{j-1}, h_j, j)$ is the transition function. Each $t_l(h_{j-1}, h_j, j)$ depends on pairs of hidden variables in the model. Index l represent all possible transitions between different hidden states.

What differentiates our LMDE model from the original work of Morency et al. is the definition of $\mathbf{S}_s(\mathbf{h}, \mathbf{x})$:

$$\mathbf{S}_s(\mathbf{h}, \mathbf{x}) = \sum_j s_s(h_j, \phi(\mathbf{x}, j)) \quad (5)$$

where

$$\phi(\mathbf{x}, j) = [q_{j_1} q_{j_2} .. q_{j_\alpha} .. q_{j_{|e|}}]. \quad (6)$$

$|e|$ is the total number of experts. Each $s_s(h_j, \phi(\mathbf{x}, j), j)$ is a state function that depends on a single hidden variable h_j and the expert output vector $\phi(\mathbf{x}, j)$. Total number of indices s is equal to the number of experts $|e|$ times the total number of hidden states $|\mathcal{H}|$. Each transition/state function is associated with a value in the corresponding model parameters (θ_l and θ_s), which can be seen as a weight assigned to this function. For each hidden state $\mathcal{H}_{y_j}$, there is a subset of $|e|$ model parameters in θ_s weighting the different expert output. Therefore, using more than one hidden states per label allows us to learn multiple mixture of experts. Each q_{j_α} is the marginal probability of expert α at frame j, and equals to $P_\alpha(y_j = a|\mathbf{x}, \lambda_\alpha)$. Each expert conditional distribution is defined by $P_\alpha(\mathbf{y}|\mathbf{x}, \lambda_\alpha)$ using the usual conditional random field formulation:

[2]This is possible because of the transition weights learned during training of the LMDE model. See Section III-B

[3]The generative model HMM with multiple states per class could also exhibit such memorization. However, our LMDE model is a discriminative model that takes into account all other labels.

$$P_\alpha(\mathbf{y}|\ \mathbf{x}, \lambda_\alpha) = \frac{\exp\left(\sum_k \lambda_{\alpha,k} \cdot \mathbf{F}_{\alpha,k}(\mathbf{y}, \mathbf{x})\right)}{\mathcal{Z}_\alpha(\mathbf{x}, \lambda_\alpha)}, \quad (7)$$

where λ_α represent the model parameters of each expert α. $\mathbf{F}_{\alpha,k}$ is defined as

$$\mathbf{F}_{\alpha,k}(\mathbf{y}, \mathbf{x}) = \sum_{j=1}^{m} f_{\alpha,k}(y_{j-1}, y_j, \mathbf{x}, j),$$

and each feature function $f_{\alpha,k}(y_{j-1}, y_j, \mathbf{x}, j)$ is either a state function $s_k(y_j, \mathbf{x}, j)$ or a transition function $t_k(y_{j-1}, y_j, \mathbf{x}, j)$. Each expert α contains a different subset of state functions $s_k(y, \mathbf{x}, j)$, defined in Section VI-C.

C. Learning Model Parameters

Given a training set consisting of n labeled sequences $(\mathbf{x_i}, \mathbf{y_i})$ for $i = 1...n$, training is done in a two step process. In the first step, we learn the model parameters, λ^*_α, for each expert α by using the following objective function from [32], [5]:

$$L(\lambda_\alpha) = \sum_{i=1}^{n} \log P_\alpha(\mathbf{y}_i \mid \mathbf{x}_i, \lambda_\alpha) - R(\lambda_\alpha) \quad (8)$$

The first term in Equation 8 is the conditional log-likelihood of the training data. The second term is a regularization term, which can be seen as assuming a prior distribution over model parameter. The two most commonly used priors are Gaussian (L_2 regularizer) and Exponential (L_1 regularizer) priors. In our experiments, we choose to use the Gaussian prior since it consistently gives better prediction results. In Section IV, this prior will be replaced by an Exponential prior in our sparse ranking algorithm.

A Gaussian prior assumes that each model parameter is drawn independently from a Gaussian distribution and penalizes according to the weighted square of the model parameters. It is defined as follows:

$$R(\lambda_\alpha) = \frac{1}{2\sigma^2}||\lambda_\alpha||^2 \quad (9)$$

where σ^2 is the variance, i.e. $P(\lambda_\alpha) \sim \exp\left(\frac{1}{2\sigma^2}||\lambda_\alpha||^2\right)$. A Gaussian prior provides smoothing when the number of learned parameters is very high compared to the size of available data. Using a Gaussian prior results in a convex quadratic optimization function that can be solved by standard optimization techniques. The marginal probabilities $P_\alpha(y_j = a \mid \mathbf{x}, \lambda^*_\alpha)$, are computed using belief propagation. In our experiments, we performed gradient ascent using the BFGS optimization technique [33].

In the second step, we use the following objective function to learn the optimal parameter θ^*:

$$L(\theta) = \sum_{i=1}^{n} \log P(\mathbf{y}_i \mid \mathbf{x}_i, \lambda_\alpha) - \frac{1}{2\sigma^2}||\theta||^2 \quad (10)$$

The first term is the conditional log-likelihood of the training data. The second term is the log of a Gaussian prior with variance σ^2.

Similar to the first step, we use gradient ascent with the BFGS optimization technique to search for the optimal parameter values, θ^*.

D. Inference

Similar to parameter learning process, inference is also achieved in two steps. Given a new test sequence $\mathbf{x}$, we first compute the marginal probabilities $P_\alpha(y_j = a|\ \mathbf{x}, \lambda_\alpha)$ for each expert. Secondly, we estimate the most probable sequence of labels $\mathbf{y}^*$ that maximizes our LMDE model:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{\mathbf{h}:\forall h_i \in \mathcal{H}_{y_i}} P(\mathbf{h} \mid \mathbf{x}, \theta^*) \quad (11)$$

where θ^* is the parameter values learned from training. To estimate the label y^*_j of frame j, we first compute the marginal probabilities $P(h_j = a|\mathbf{x}, \theta^*)$ for all possible hidden states H. Then, we sum the marginal probabilities according to the disjoint sets of hidden states H_{y_j}. Finally, the label y^*_j associated with the optimal set is chosen.

IV. Sparse Feature Ranking

One advantage of the LMDE model is that it can be easily interpreted to see what was learned. In this section, we present a feature ranking scheme that allows us to find the subset of features that are the most relevant to each expert.

Our feature ranking scheme relies on sparse regularization of LMDE model parameters λ_α. Using a regularization term in the optimization function during training can be seen as assuming a prior distribution over the model parameter. In Section III, a Gaussian prior(L_2-regularizer) was preferred due

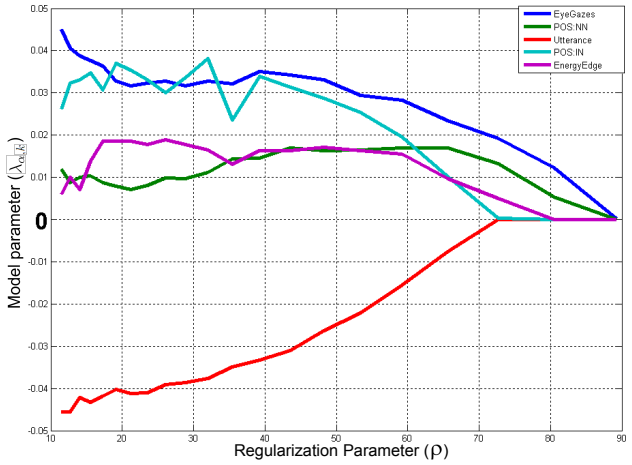


Fig. 4. **Example of sparse ranking using L_1 regularization.** As ρ goes from higher to lower values, model parameters start to become non-zero based on their relevance to the prediction model.

to better classification performance. But here, we replace the L_2-regularization term of Equation 8 by an Exponential prior (L_1-regularizer), which allows us to better analyze the model. L_1 regularization results in sparse parameters vector in which many of the parameters are exactly zero [34]. Therefore, it has been widely used in different domains for the purpose of feature selection [35], [3]. An Exponential prior penalizes according to the weighted L_1 norm of the parameters and is defined as follows:

$$R(\lambda_\alpha) = \rho \parallel \lambda_\alpha \parallel_1 \tag{12}$$

where $\rho > 0$ and determines how much penalty should be applied by the regularization term. Larger values indicate larger penalty, thus producing sparser parameter vector λ_α.

Figure 4 shows an example of how ρ effects the model parameters. In this example, we trained a single expert with 5 input features: EyeGazes, POS:NN, utterance, POS:IN, and EnergyEdge (see Section VI-C for details about feature representations). Figure 4 shows the effect of regularization on model parameters λ_α. This regularization path was created by starting with a high regularization penalty ρ where all the parameters are zero and then gradually reduce the regularization until all the parameters have non-zero values. In this path, if a parameter becomes non-zero in earlier stages (i.e., large ρ), this signifies the input feature associated with this parameter is important. Our ranking scheme is based on this observation. We rank the features in the order of them becoming non-zero in the regularization path. For the example shown in Figure 4, our algorithm will rank the features as follows: (1) EyeGazes and POS:NN, (2) EnergyEdge, (3) Utterance and POS:IN. The pseudo code for our sparse feature ranking approach is given in Algorithm 1.

Algorithm 1 Sparse Feature Ranking

```
ranked_features = empty
for ρ = ∞ down to 0 do
    train an expert CRF with L-1 regularization factor ρ
    for all nonzero feature params λ_{α,k} do
        if λ_{α,k} is NOT in selected_features then
            ranked_features = {ranked_features, f_{i,k}}
        end if
    end for
end for
return ranked_features
```

The regularization penalty ρ determines how sparse the model should be. More than one of these parameters may become non-zero at any given ρ regularization factor. Therefore, our feature ranking scheme allows more than one feature to have the same rank, meaning that these features have equivalent influence and they should be selected together. Compared to other greedy methods [29], our sparse feature ranking algorithm is non-greedy in the sense that all features are present during selection process. Also, our algorithm is much more efficient than the greedy method, since the computational cost of our algorithm is determined by the number of regularization penalty values ρ used (which was 76 in our experiments). On the other hand, the computational cost of the greedy approach increases with the number of features (a total of 1629 features are used in our experiments).

Using an L_1-regularizer results non-differentiable objective function. Therefore, we use Orthant-Wise Limited-memory Quasi-Newton (owl-qn) method [36] for training L1-regularized log-linear models, which is an extension of L-BFGS optimization technique.

V. LMDE for Multimodal Prediction

LMDE is a generic approach designed to integrate information from multiple modalities. In this

section, we first provide a detailed discussion about multimodal prediction, and more specifically about backchannel prediction which is used as the main task in our experiments. Then, we present the User-adaptive Prediction Accuracy, a new evaluation metric for prediction models.

A. Multimodal Prediction

Human face-to-face communication is a little like a dance, in that participants continuously adjust their behaviors based on verbal and nonverbal displays and signals. A topic of central interest in modeling such behaviors is the patterning of interlocutor actions and interactions, moment-by-moment, and one of the key challenges is identifying the patterns that best predict specific actions. Thus we are interested in developing predictive models of communication dynamics that integrate previous and current actions from all interlocutors to anticipate the most likely next actions of one or all interlocutors. Humans are good at this: they have an amazing ability to predict, at a micro-level, the actions of an interlocutor [37]; and we know that better predictions can correlate with more empathy and better outcomes [38], [39].

Building computational models of such a predictive process involves dynamics and temporal relationship between cues from different modalities [40]. These different modalities contain complementary information essential to interpretation and understanding of human behaviors [41]. Psycholinguistic studies also suggest that gesture and speech come from a single underlying mental process, and they are related both temporally and semantically [42], [43], [44].

Among other behaviors, backchannel feedback (the nods and paraverbals such as "uh-hu" and "mm-hmm" that listeners produce as some is speaking) has received considerable interest due to its pervasiveness across languages and conversational contexts. Several systems have been demonstrated on the task of listener backchannel feedback prediction [8], [27], [29]. Evaluation of results from a backchannel prediction model is challenging, since listener feedback varies between people and is often optional. While experiencing the same set of environmental conditions, some people may choose to give more frequent feedbacks, whereas some others may choose to be less active and give seldom feedbacks. Therefore, results from prediction tasks

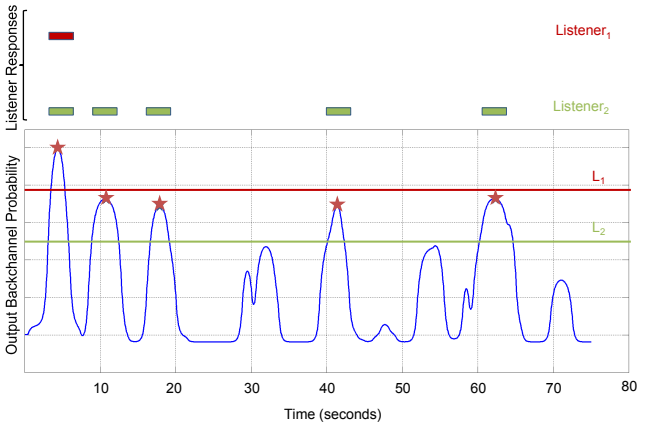


Fig. 5. A sample output sequence of listener feedback probabilities (in blue). Red and green boxes indicate the responses from $Listener_1$ and $Listener_2$ respectively. The red and green lines indicate the thresholds on the output probabilities that can correctly assign the backchannel labels to the corresponding listener labels.

are expected to have lower accuracies as opposed to recognition tasks where the data labels are well established. This indicates the necessity of a new error measurement, which can take into account differences in human behaviors. We address this issue in the next section.

B. User-adaptive Prediction Accuracy

The traditional way to evaluate prediction models is usually to set a threshold on the output probability, so that final decision can be made (i.e. backchannel or not). From these final predictions, typical error metrics, such as F1-score, precision and recall, can be measured. The same threshold will be applied to all data sequences from different people in the test set. However, people do not always respond the same way to the same stimuli (e.g. speaker's actions). Some people may naturally give a lot of feedback while others will give feedback only when the speaker is directly requesting it. For this reason, using the same threshold for evaluating multiple listeners may not be representative of the real predictive power of the learned model (e.g. LMDE).

Let's illustrate this problem with an example as depicted in Figure 5. In this case, we have two listeners listening to the same speaker, but reacting differently. $Listener_1$ gave only 1 backchannel feedback, while $Listener_2$ was more actively nodding his head and gave 5 backchannels. Figure 5 shows the output of our LMDE model (backchannel

probabilities) as a continuous blue line and the potential predictions (local maxima) are depicted by the red stars. The question now is: can our learned model correctly predict both listeners? As shown in the figure, there is not one threshold that can correctly predict both listeners' behaviors. However, given the right thresholds, this model can correctly predict both listeners. So, what should be the evaluation measure and the performance of the model?

To address this issue, we propose a new error measurement called User-adaptive Prediction Accuracy (UPA). The main intuition behind UPA is that we will ask our prediction model to give us the n_i-best predictions, where n_i is the number of times that a particular $listener_i$ gave a backchannel. Following this intuition, UPA is defined as:

$$UPA = \frac{1}{L}\sum_{i}^{N}\frac{P(n_i)}{n_i/l_i} \tag{13}$$

where i is the listener id, N is the total number of listeners in the test data, n_i is the number of backchannels $listener_i$ provided during a dyadic interaction, and l_i is the length of the interaction i. Therefore, the denominator term coveys the backchannel frequency of $listener_i$. L is the total length of all interactions with all the listeners. $P(n_i)$ is a function that compares the n_i-best predictions from our LMDE model output to the ground truth backchannel labels from $listener_i$. The function $P(n_i)$ returns the number of correctly predicted listener backchannels. Predictions from our LMDE model are ranked by their probability output.

UPA gives us a measure of the prediction quality while adapting to people's different levels of backchannel responses. Consider the case where two different listeners gave the same amount of backchannel during their interactions, and the duration of first interaction with one of the listeners is much longer than the duration of second interaction with the other listener. One would expect more noise (i.e. peaks) in the output probabilities of the first interaction corresponding to possible backchannel opportunities that the actual listener had missed. Therefore, a model that can correctly find the true backchannel opportunities even if the listener rarely provides backchannel should be given a higher weight. Therefore, we introduce l_i weighting in Equation 13 to capture these differences in listener's responses. Our UPA performance measure, by removing the performance variance due to the variability in amount of backchanneling, is a more reliable performance measure than standard measures like precision, recall and F-measure.

VI. Experimental Setup

As mentioned in the previous section, we evaluate our LMDE on the multimodal task of predicting listener nonverbal backchannel. In this section, we first describe our dataset, backchannel annotation technique and multimodal speaker features. Then, we explain the baseline models used for comparison in our tests, and the experimental setup.

A. Dataset

We are using the RAPPORT dataset [4] from [4], which contains 45 dyadic interactions between a speaker and a listener. Data is drawn from a study of face-to-face narrative discourse ("quasi-monologic" storytelling). In this dataset, participants in groups of two were told they were participating in a study to evaluate a communicative technology. Subjects were randomly assigned the role of speaker and listener. The speaker viewed a short segment of a video clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video. After the speaker finished viewing the video, the listener was led back into the computer room, where the speaker was instructed to retell the stories portrayed in the clips to the listener. The listener was asked to not talk during the story retelling. Elicited stories were approximately two minutes in length on average. Participants sat approximately 8 feet apart. All video sequences were manually transcribed and manually annotated to determine the ground truth backchannels. The next section describes our annotation procedure.

B. Backchannel Annotations

In our experiments, we focus on visual backchannels: head nods. A head nod gesture starts when the person starts moving his/her head vertically.

[4]This dataset has also been used in [29] in which 50 interactions were reported. In that study, the only visual cue exploited is the speaker eye gaze. We have extracted more visual cues (smiles, eye brows) from the speaker videos and omitted the sequences for which the speaker videos were not completely annotated.

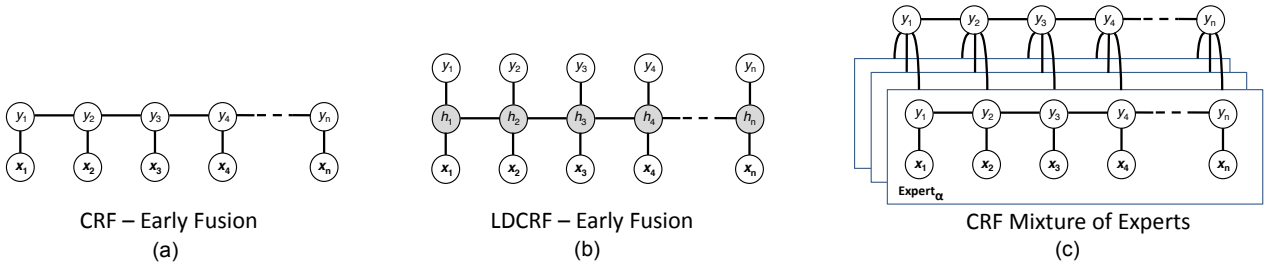


Fig. 6. Baseline Models: **a)** Conditional Random Fields (CRF), **b)** Latent Dynamic Conditional Random Fields(LDCRF), **c)** CRF Mixture of Experts (no latent variable)

The head nod gesture ends when the person stops moving or when a new head nod is started. A new head nod starts if the amplitude of the current head cycle is higher than the previous head cycle. Some listeners' responses may be longer than others although they all correspond to one single respond. In our data, annotators found a total of 666 head nods. The duration of these nods varied from 0.16 seconds to 7.73 seconds. Mean and standard deviation of backchannel durations are 1.6 and 1.2 respectively. The minimum number of head nods given by one listener during one interaction is 1, the maximum is 47, mean and standard deviations are 14.8 and 10.9 respectively.

Following Ward and Tsukahara's [8] original work on backchannel prediction, we train our LMDE model to predict only the start time of the backchannel start cue (i.e. head nod). Following again Ward and Tsukahara [8], we define the backchannel duration as a window of 1.0 seconds centered around the start time of the backchannel. A backchannel cue will be correctly predicted if at least one prediction of our LMDE model happens during this 1.0 seconds duration. All models tested in this paper use this same testing backchannel duration of 1.0 seconds. During the training of our LMDE prediction model, we will vary the backchannel duration to see which one is optimal. The Section VII-B describes these results, where we find the optimal training backchannel duration to be 0.5 seconds.

C. Multimodal Features and Experts

This section describes the different multimodal features used to create our five experts.

PROSODY Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody [45], [8], [46]. For example, Ward and Tsukahara [8] show that short listener backchannels (listener utterances like "ok" or "uh-huh" given during a speaker's utterance) are associated with a lowering of pitch over some interval. Listener feedback often follows speaker pauses or filled pauses such as "um" (see [46]). Using openSMILE [47] toolbox, we extract the following prosodic features, including standard linguistic annotations and the prosodic features suggested by Ward and Tsukhara:

- downslopes in pitch continuing for at least 40ms
- regions of pitch lower than the 26th percentile continuing for at least 110ms (i.e., lowness)
- drop or rise in energy of speech (i.e., energy edge)
- fast drop or rise in energy of speech (i.e., energy fast edge)
- vowel volume (i.e., vowels are usually spoken softer)
- pause in speech (i.e., no speech)

VISUAL GESTURES Gestures performed by the speaker are often correlated with listener feedback [48]. Eye gaze, in particular, has often been implicated as eliciting listener feedback. Thus, we manually annotate the following contextual features:

- speaker looking at listener (eye gaze)
- speaker not looking at listener (˜ eye gaze)
- smiling
- moving eyebrows up
- moving eyebrows down

LEXICAL Some studies have suggested an association between lexical features and listener feedback [46]. Using the transcriptions, we included all individual words (i.e., unigrams) spoken by the speaker

during the interactions.

PART-OF-SPEECH TAGS In [46], combination of pause duration and a statistical part-of-speech language model is shown to achieve the best performance for placing backchannels. Following this work, we use a CRF part-of-speech (POS) tagger to automatically assign a part of speech label to each word. We also include these part-of-speech tags (e.g. noun, verb, etc.) in our experiments.

SYNTACTIC STRUCTURE Finally, we attempt to capture syntactic information that may provide relevant cues by extracting three types of features from a syntactic dependency structure corresponding to the utterance. The syntactic structure is produced automatically using a data-driven left-to-right shift-reduce dependency parser [49], trained POS on dependency trees extracted from the Switchboard section of the Penn Treebank [50], converted to dependency trees using the Penn2Malt tool[5]. The three syntactic features are:

- Grammatical function for each word (e.g. subject, object, etc.), taken directly from the dependency labels produced by the parser
- Part-of-speech of the syntactic head of each word, taken from the dependency links produced by the parser
- Distance and direction from each word to its syntactic head, computed from the dependency links produced by the parser

Although our current method for extracting these features requires that the entire utterance be available for processing, this provides us with a first step towards integrating information about syntactic structure in multimodal prediction models. Many of these features could in principle be computed incrementally with only a slight degradation in accuracy, with the exception of features that require dependency links where a word's syntactic head is to the right of the word itself. We leave an investigation that examines only syntactic features that can be produced incrementally in real time as future work.

D. Baseline Models

INDIVIDUAL EXPERTS Our first baseline model consists of a set of CRF chain models, each trained with different set of multimodal features (as described in the previous section). In other words, only visual, prosodic, lexical or syntactic features are used to train a single CRF expert. (See Figure 6a).

MULTIMODAL CLASSIFIERS (EARLY FUSION) Our second baseline consists of two models: CRF and LDCRF [6]. To train these models, we concatenate all multimodal features (lexical, syntactic, prosodic and visual) in one input vector. Graphical representation of these baseline models are given in Figure 6-(a) and Figure 6-(b).

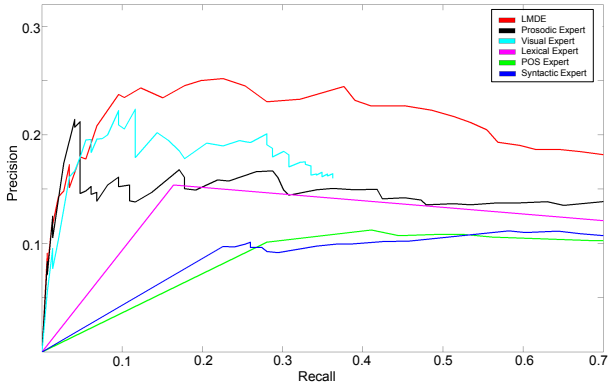
CRF MIXTURE OF EXPERTS To show the importance of latent variable in our LMDE model, we trained a CRF-based mixture of discriminative experts. A graphical representation of a CRF Mixture of experts is given in Figure 6. This model is similar to the Logarithmic Opinion Pool (LOP) CRF suggested by Smith et al. [7], in the sense that they both factor the CRF distribution into a weighted product of individual expert CRF distributions. However, the main difference between LOP and CRF Mixture of Experts model is in the definition of optimization functions. Similar to our LMDE model, training of CRF Mixture of Experts is performed in two steps: Expert models are learned in the first step, and the second level CRF model parameters are learned in the second step.

PAUSE-RANDOM CLASSIFIER Our last baseline model is a random backchannel generator, which randomly generates backchannels whenever some pre-defined conditions in the speech is purveyed. These conditions include pauses that come after at least 700 milliseconds of speech and absence of backchannel feedback within the preceeding 800 milliseconds. This random classifier has also been used by Ward and Tsukahara [8] for comparison.

E. Methodology

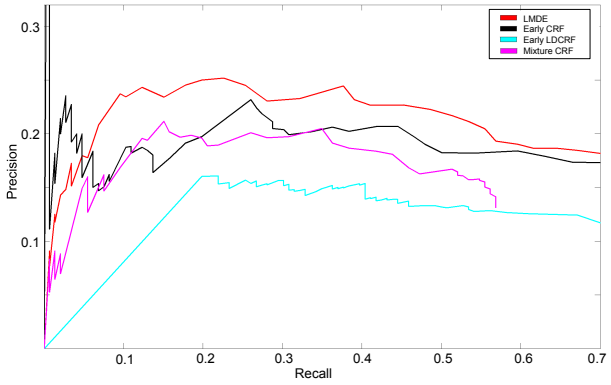
We performed held-out testing by randomly selecting a subset of 11 interactions (out of 45) for the test set. The training set contains the remaining 34 dyadic interactions. All models in this paper were evaluated with the same training and test sets. Validation of all model parameters (regularization term and number of hidden states) was performed using a

[5]http://w3.msi.vxu.se/ nivre/research/Penn2Malt.html



Expert	Precision	Recall	f1	upa
LMDE	**0.1984**	**0.6369**	**0.3026**	**0.2640**
Prosodic	0.1423	0.5188	0.2233	0.1848
Visual	0.1462	0.2768	0.1914	0.1558
Lexical	0.1171	0.5701	0.1943	0.1131
POS	0.1247	0.3705	0.1866	0.1122
Syntactic	0.0954	0.2595	0.1395	0.1252

Fig. 7. Comparison of individual experts with our LMDE model. **Top:** Recall (x-axis) v.s. Precision (y-axis) values for different threshold values. **Bottom:** Precision, Recall, F1 and UPA scores of corresponding models for selected amount of backchannel.



model	Precision	Recall	f1	gpe
LMDE	**0.1984**	**0.6369**	**0.3026**	**0.2640**
Early CRF	0.1724	0.4630	0.2512	0.1615
Early LDCRF	0.1240	0.2414	0.1638	0.0788
Mixture CRF	0.1607	0.4984	0.2430	0.2027
Pause-Random	0.0843	0.2945	0.1306	0.0883

Fig. 8. Comparison of our LMDE model with previously published approaches for multimodal prediction. **Top:** Recall (x-axis) v.s. Precision (y-axis) values for different threshold values. **Bottom:** Precision, Recall, F1 and UPA scores of corresponding models for selected amount of backchannel.

3-fold cross-validation strategy on the training set. The regularization term was validated with values $10^k, k = -1..3$. Two different number of hidden states were tested for the LMDE models: 2, and 3 (note that LMDE with 1 hidden state is equivalent to Mixture of CRF Experts model). In our experiments, the optimum number of hidden states was 2 when duration of backchannel labels was set to 0.5, and 3 when duration of backchannel labels was set to 1.0 or 1.5.

The performance is measured by using UPA (described in Section V-B) as well as more conventional metrics: precision, recall, and F-measure. Precision is the probability that predicted backchannels correspond to actual listener behavior. Recall is the probability that a backchannel produced by a listener in our test set was predicted by the model. We use the same weight for both precision and recall, so-called F_1, which is the weighted harmonic mean of precision and recall. F_1 scores for each sequence is calculated first, then the final F_1 result is computed by averaging these sequence scores.

During testing, we find all the "peaks" (i.e., local maxima) from marginal probabilities $P(y_j = a|\ \mathbf{x}, \theta)$. When computing UPA, the final predictions are selected from these peaks so that the number of model predictions are equal to the number of listener backchannels in the test sequence. For the f1-score, the prediction model needs to decide on a specific threshold (i.e., amount of backchannel) for the marginal probabilities for all users. The value of this threshold is automatically set during validation. Since we are predicting the start time of a backchannel, an actual listener backchannel is correctly predicted if at least one model prediction happen within the 1 second interval window around the start time of the listener backchannel.

The training of all CRFs and LDCRFs were done using the hCRF library[6]. The LMDE model was implemented in Matlab based on the hCRF library. The input observations were computed at 30 frames per second. Given the continuous labeling nature of our LMDE model, prediction outputs were also computed at 30Hz.

VII. Results

In this section we present the results of our empirical evaluation. We designed our experiments so to test different characteristics of the LMDE model. First, we present our quantitative results that evaluate: (1) integration of multiple sources of information, (2) late fusion approach and (3) latent

[6]http://sourceforge.net/projects/hrcf/

variable which models the hidden dynamic between experts. Then, we present qualitative analysis related to: (1) the output probabilities from individual experts and the LMDE model, (2) the most relevant features in early and late fusion models, (3) model robustness and (4) UPA analysis.

A. Comparative Results

INDIVIDUAL EXPERTS We trained one individual expert for each feature types: visual, prosodic, lexical and syntactic features (both part-of speech and syntactic structure). Precision, recall, F_1, and UPA values for each individual expert and our LMDE model are shown in Figure 7 (Bottom) [7]. Even though the experts may not perform well individually, they can bring important information once merged together. Recall-precision curve in Figure 7 (Top) shows that our LMDE model was able to take advantage of the complementary information from each expert.

LATE FUSION We compare our approach with two early fusion models: CRF and LDCRF (see Figure 6). Figure 8 summarize the results. The CRF model learns direct weights between input features and the gesture labels. The LDCRF is able to model more complex dynamics between input features with the latent variable. We can see that our LMDE model outperforms both early fusion model because of its late fusion approach.

When merging the features together in an early manner, the noise from one modality may hide or suppress the features from a different modality. By training separate experts for each different modality, we are able to reduce the effect of this noise, therefore learn models that can generalize better to new multimodal data.

LATENT VARIABLE The CRF Mixture of Experts [7] directly merges the expert outputs while our LMDE model uses a latent variable to model the hidden dynamic between experts (see Figure 6-(c)). This comparison (summarized in Figure 8) is important since it shows the effect of the latent variable in our LMDE model.

[7]While calculating these values, we first find the peaks from marginal probabilities for possible prediction points. However, there is no guarantee that a peak will appear during all ground truth backchannel regions. Therefore, we cannot get higher recall values for some of the experts, i.e. Visual Expert.

TABLE II
PERFORMANCES OF INDIVIDUAL EXPERT MODELS TRAINED BY USING ONLY THE TOP 5 FEATURES SELECTED BY OUR FEATURE RANKING ALGORITHM. THE LAST TWO ROWS REPRESENT THE LMDE MODELS USING THE EXPERT MODELS TRAINED WITH ONLY 5 FEATURES SELECTED BY EITHER BY A GREEDY METHOD [29] OR OUR SPARE FEATURE RANKING SCHEME.

Expert	Precision	Recall	f1	upa
Prosodic5	0.1463	0.5645	0.2324	0.1545
Visual5	0.1457	0.2671	0.1886	0.1558
Lexical5	0.1059	0.1706	0.1307	0.1471
POS5	0.1522	0.5602	0.2394	0.1409
Syntactic5	0.0995	0.5626	0.1691	0.1302
Greedy5	0.2007	0.3241	0.2479	0.2585
LMDE5	0.1914	0.5306	0.2814	0.2331

B. Analysis Results

EXPERT ANALYSIS Our first analysis looks at speaker features that are the most relevant to listener feedback prediction. This analysis is performed by applying our sparse feature ranking algorithm described in Section IV to each expert separately. Top 5 features for our five experts are listed in Table I [8]. First interesting results are the two features appearing in Prosodic Expert and one feature appearing in Visual Expert: *pause*, *low pitch* and *eye gaze*. These features have also been identified in previous work [8], [29] as important cues for backchannelling. Similarly, *um* feature in Lexical Expert can be considered as a filler *pause* and reasonable cue for backchannel prediction. Visual Expert selects *nod* as the second-best feature, which can be associated with mirroring effect. This suggests that our experts are learning relevant features.

To confirm that these selected features are relevant to the L2 trained models, we trained new experts and LMDE models using the top 5 features of each expert selected by our sparse feature ranking algorithm. In other words, for each expert, we trained a new CRF model by using only the top 5 features selected for that expert. Performance of these new expert models are listed in Table II. It is

[8]Utterence indicates when the user is talking. POS:NN indicates singular noun, POS:PRP indicates personal pronoun, POS:VBG indicates verb, POS:UH indicates interjection and POS:NNS indicates plural noun. DIRDIST:L1 and L2 describe the distance and direction from the head node in the parse tree (i.e. left within distance 1). LABEL:PMOD and LABEL:SUB indicate a proposition and a subject modifier respectively. HEADPOS:VBZ indicatea verb head node.

TABLE I
TOP 5 FEATURES FROM RANKED LIST OF FEATURES FOR EACH LISTENER EXPERT.

Expert 1 (Prosodic)	**Expert 2 (Visual)**	**Expert 3 (Lexical)**	**Expert 4 (POS)**	**Expert 5 (Syntactic)**
Utterence	˜ EyeGaze	she	POS:NN	DIRDIST:L1
Pause	Nod	um	POS:PRP	HEADPOS:VBZ
Vowel Volume	EyeBrows Up	that	POS:VBG	LABEL:PMOD
Energy Edge	EyeBrows Down	he	POS:UH	DIRDIST:L2
Low Pitch	EyeGaze	women	POS:NNS	LABEL:SUB

interesting to see that using only five features can achieve performance as good as when using all the features. We see some increase in both f1 and upa values for POS and Syntactic experts when 5 features used. We believe that this is due to noise when all features are used. For comparison, we trained a new LMDE model using these new expert models. The performance of this model, which we refer to as LMDE5, is given in Table II. LMDE5 achieves a higher f1 value than all individual experts, and a very similar upa value as the original LMDE (remark that all the features were present while training the expert models in the original LMDE).

We also compared our sparse feature ranking algorithm to the greedy feature selection method presented in [29]. For this purpose, we used this greedy method to select 5 features for each expert, and learned expert models trained with these 5 features. Then, these expert models are used to learn an LMDE model, referred to as Greedy5. The results are shown in Table II. LMDE5 and Greedy5 achieved similar performance. However, our sparse ranking scheme is a much faster algorithm than the greedy method. The computational cost of the greedy algorithm increases with the number of features, whereas the computational cost of our spare ranking scheme is determined by the number of regularization penalty values ρ (76 in our experiments).

LMDE MODEL ANALYSIS Our second analysis focuses on the multimodal integration which happens at the latent variable level in our LMDE model, Figure 9 shows the output probabilities for all five individual experts as well as our model. The strength of the latent variable is to enable different weighting of the experts at different point in time.

In the sequence depicted in Figure 9, the actual

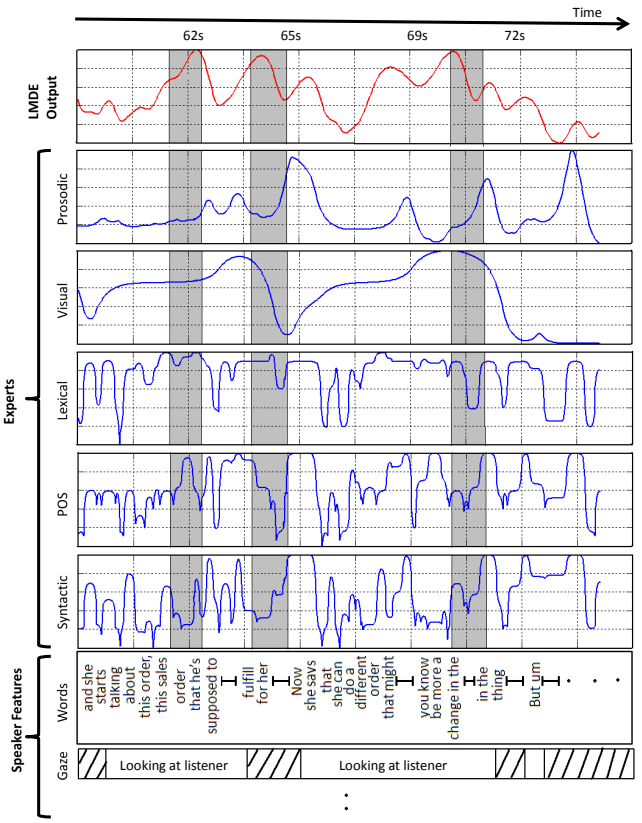


Fig. 9. Output probabilities from LMDE and individual experts for two different sub-sequences. The gray areas in the graph correspond to ground truth backchannel feedbacks of the listener.

listener gave backchannel feedback 3 times (around 62s, 64s and 71s), which are indicated by the gray areas. As we analyze the outputs from different experts, we see that the Lexical and POS experts were able to learn the backchannel opportunity for the first backchannel feedback at 62s. These two experts are highly weighted (by one of the hidden state) during this part of the sequence. All the experts except the Visual Experts assigned a high chance of backchannel around 65.5s, where there is no listener feedback. The Visual Expert was highly

TABLE IV
PERFORMACES OF BASELINE MODELS AND OUR LMDE MODEL AS WE INCREASE THE DURATION OF BACKCHANNEL LABELS DURING TRAINING.

	Training Backchannel Duration					
	0.5		**1.0**		**1.5**	
	f1	upa	f1	upa	f1	upa
LMDE	**0.3026**	**0.2640**	**0.2774**	**0.2439**	**0.2751**	**0.2291**
Early CRF	0.2512	0.1615	0.2384	0.1916	0.2397	0.1764
Early LDCRF	0.1638	0.0788	0.1856	0.0669	0.1648	0.0496
Mixture CRF	0.2430	0.2027	0.2245	0.1834	0.2037	0.1576

TABLE III
NUMBER OF BACKCHANNEL FEEDBACKS PROVIDED BY EACH OF THE 11 LISTENERS IN OUR TEST SET AND THEIR CORRESPONDING UPA, PRECISION, RECALL AND F1 SCORE.

num of feedbacks	upa	Precision	Recall	f1
1	0.000	0.031	1.000	0.061
1	0.000	0.050	1.000	0.095
2	0.000	0.031	0.500	0.059
4	0.000	0.077	0.500	0.133
5	0.200	0.091	0.600	0.158
8	0.375	0.104	0.625	0.178
16	0.562	0.282	0.687	0.400
21	0.238	0.269	0.333	0.298
23	0.478	0.433	0.565	0.491
25	0.320	0.286	0.720	0.409
40	0.500	0.528	0.475	0.500

weighted during this time, so that the influence of all other experts was reduced in the LMDE output. This difference of weighting shows that a different hidden state is active during this part of the sequence.

MODEL ROBUSTNESS As mentioned in Section VI-B, one of the hyper-parameter of our LMDE prediction model is the duration of backchannel cues used during training. To analyze sensitivity of our model to backchannel duration, we varied the duration from 500 seconds to 1500 seconds, and retrained our LMDE model and the baseline models. F1 and UPA values are given in Table IV. We observe a drop in the LMDE performance as we increase the duration. This was true for most of the other models, which suggests that it is better to train prediction models with more focused labels (i.e. narrow backchannel duration). It should also be noted that LMDE outperforms all other baseline models for all different durations.

UPA ANALYSIS

In our earlier experiments (see Figure 7, we have seen that the Visual and Lexical experts seem to perform about the same based on their F_1 values (0.1914 and 0.1943), but their UPA values are quite different (0.1558 and 0.1131). Looking at their F_1 results, we would expect these two experts to have very similar Recall-Precision curves. However, their recall-precision curves in Figure 7 indicate that the Visual Expert is a better model than the Lexical Expert, which is already confirmed by our UPA measure. We can see another such example between the POS and Syntactic Experts. The F_1 values indicate that POS Expert (0.1866) is a better model than the Syntactic Expert(0.1395). On the other hand, their UPA values(0.1122 and 0.1252) tell that these are similar models, which is also confirmed by their Precision-Recall curves in Figure 7. These observations suggest that our UPA measurement is a more representative measure than the F_1 score.

To analyze the variability among listeners, we have listed in Table III the individual test performances and the number of backchannel feedback provided by each listener. One interesting conclusion derived from this result is that there is some correlation with the number of feedbacks and upa, precision, and f1 values. As the number of backchannels increase, these values increase as well.

VIII. CONCLUSION

In this paper, we addressed three main issues involved in building predictive models of human communicative behaviors. First, we introduced a new model called Latent Mixture of Discriminative Experts (LMDE) for multimodal data integration.

Many of the interactions between speech and gesture happen at the sub-gesture or sub-word level. LMDE learns automatically the temporal relationship between different modalities. Since, we train separate experts for each modality, LMDE is capable of improving the prediction performance even with limited amount of data.

We evaluated our model on the task of nonverbal feedback prediction (e.g., head nod). Our experiments confirm the importance of combining the five types of multimodal features: lexical, syntactic structure, POS, visual, and prosody. An important advantage of using our LMDE model is that it enables easy interpretability of individual experts. As a second contribution, we have presented a sparse feature ranking scheme based on L_1 regularization technique. Our third contribution is a new metric called User-adaptive Prediction Accuracy (UPA). This metric is particularly designed for evaluating prediction models, and we plan to apply it to other prediction models as well. LMDE is a generic model that can be applied to a wide range of problems. In the future, we want to apply it for other multimodal prediction tasks.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0917321 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] A. Y. Ng, "Feature selection, l-1 vs. l-2 regularization, and rotational invariance," in *International Conference on Machine Learning (ICML)*, 2004.

[2] A. Smith and M. Osborne, "Regularisation techniques for conditional random fields: Parameterised versus parameter-free," in *International Joint Conference on Natural Language Processing (NLP)*, 2005.

[3] D. L. Vail, "Feature selection in conditional random fields for activity recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007.

[4] J. Gratch, N. Wang, J. Gerten, and E. Fast, "Creating rapport with virtual agents," in *Intelligent Virtual Agents (IVA)*, 2007.

[5] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labelling sequence data," in *International Conference on Machine Learning (ICML)*, 2001.

[6] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[7] A. Smith, T. Cohn, and M. Osborne, "Logarithmic opinion pools for conditional random fields," in *Association for Computational Linguistics (ACL)*, 2005, pp. 18–25.

[8] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *Journal of Pragmatics*, vol. 23, pp. 1177–1207, 2000.

[9] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[10] V. Pavlovic, "Multimodal tracking and classification of audio-visual features," in *IEEE International Conference on Image Processing (ICIP)*, vol. I, 1998, pp. 343–347.

[11] L. Terry, D. Shiell, and A. Katsaggelos, "Feature space video stream consistency estimation for dynamic stream weighting in audio-visual speech recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2008, pp. 1316–1319.

[12] N. Fox and R. Reilly, "Audio-visual speaker identification based on the use of dynamic audio and visual features," in *International Conference on Audio and Video-based Biometric Person Authentication (IAPR)*, 2003, pp. 743–751.

[13] I. A. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, March 2005.

[14] S. Foo, Y. Lian, and L. Dong, "Recognition of visual speech elements using adaptively boosted hidden markov models," in *IEEE Transactions on Circuits and Systems for Video Technology*, 2004, pp. 693–705.

[15] A. Garg, V. Pavlovic, and J. Rehg, "Boosted learning in dynamic bayesian networks for multimodal speaker detection," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1355–1369, September 2003.

[16] J. Peng, L. Bo, and J. Xu, "Conditional neural fields," in *Advances in Neural Information Processing Systems*, 2009, pp. 1419–1427.

[17] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," *ACM international conference on Multimedia*, 2005.

[18] Z. Ghahramani, M. I. Jordan, and P. Smyth, "Factorial hidden markov models," in *Machine Learning*. MIT Press, 1997.

[19] N. Oliver, A. Garg, and E. Horvitz, "Layered representations for learning and inferring office activity from multiple sensory channels," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 163–180, 2004.

[20] M. Barnard and J.-M. Odobez, "Sports event recognition using layered hmms," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005, pp. 1150–1153.

[21] M. I. Jordan, "Hierarchical mixtures of experts and the em algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[22] C. Bishop and M. Svensén, "Bayesian hierarchical mixtures of experts," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.

[23] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3d human motion estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 390–397.

[24] Nakano, Reinstein, Stocky, and J. Cassell, "Towards a model of face-to-face grounding," in *Association for Computational Linguistics (ACL)*, 2003.

[25] Y. Nakano, K. Murata, M. Enomoto, Y. Arimoto, Y. Asa, and H. Sagawa, "Predicting evidence of understanding by moni-

toring user's task manipulation in multimodal conversations," in *Association for Computational Linguistics (ACL)*, 2007, pp. 121–124.

[26] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi, "A conversation robot using head gesture recognition as para-linguistic information," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2004, pp. 159–164.

[27] M. Maatman, J. Gratch, and S. Marsella, "Natural behavior of a listening agent," in *Intelligent Virtual Agents (IVA)*, 2005.

[28] S.-H. Kang, J. Gratch, N. Wang, and J. Watt, "Does the contingency of agents' nonverbal feedback affect users' social anxiety?" in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2008.

[29] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in *Conference on Intelligent Virutal Agents (IVA)*, 2008.

[30] S. Perkins, K. Lacker, J. Theiler, I. Guyon, and A. Elisseeff, "Grafting: Fast, incremental feature selection by gradient descent in function space," *Journal of Machine Learning Research. 3, 1333–1356*, 2003.

[31] S. Riezler and A. Vasserman, "Incremental feature selection and l1 regularization for relaxed maximum-entropy modeling," in *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2004.

[32] S. Kumar and M. Herbert., "Discriminative random fields: A framework for contextual interaction in classification," in *International Conference on Computer Vision (ICCV)*, 2003.

[33] J. Nocedal and S. J. Wright, "Numerical optimizations," in *Springer Series in Operations Research*, 2006.

[34] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B. 58, 267–288*, 1994.

[35] A. McCallum and C. R, "Efficiently inducing features of conditional random fields," in *Conference on Uncertainty in Artificial Intelligence*, 2003.

[36] G. Andrew and J. Gao, "Scalable training of l1-regularized log-linear models," in *International Conference on Machine Learning (ICML)*, ser. ICML '07. ACM, 2007.

[37] J. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of Personality and Social Psychology (JPSP)*, vol. 79, no. 6, pp. 941–952, 2000.

[38] S. Goldberg, "The secrets of successful mediators," *Negotiation Journal*, vol. 21, no. 3, pp. 365–376, 2005.

[39] D. Fuchs, "Examiner familiarity effects on test performance: implications for training and practice," *Topics in Early Childhood Special Education*, vol. 7, pp. 90–104, 1987.

[40] F. Quek, "The catchment feature model for multimodal language analysis," in *International Conference on Computer Vision (ICCV)*, 2003.

[41] S. Oviatt, "Ten myths of multimodal interaction," *Communications of the ACM*, 1999.

[42] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. Univ. Chicago Press, 1992.

[43] J. Cassell and M. Stone, "Living hand to mouth: Psychological theories about speech and gesture in interactive dialogue systems," in *Conference on Artificial Intelligence (AAAI)*, 1999.

[44] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[45] R. Nishimura, N. Kitaoka, and S. Nakagawa, "A spoken dialog system for chat-like conversations considering response timing," *LNCS*, vol. 4629, pp. 599–606, 2007.

[46] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *European chapter of the Association for Computational Linguistics (EACL)*, 2003, pp. 51–58.

[47] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Affective Computing and Intelligent Interaction (ACII)*, 2009, pp. 576–581.

[48] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge: Cambridge University Press, 1995.

[49] K. Sagae and J. Tsujii, "Dependency parsing and domain adaptation with LR models and parser ensembles," in *Association for Computational Linguistics (ACL)*, 2007, pp. 1044–1050.

[50] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger, "The penn treebank: annotating predicate argument structure," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 1994, pp. 114–119.

Derya Ozkan has received her BSc, MSc degrees from Department of Computer Engineering at Bilkent University. During her master thesis she worked on recognition and retrieval of faces under the supervision of Dr. Pinar Duygulu. Between June 2006 and February 2007, she worked as a research intern at Siemens Corporate Research, Princeton, NJ. Derya is now a PhD Student at Computer Science Department, University of Southern California (USC) since August 2007. Her supervisor is Louis-Philippe Morency, and she is a member of Multimodal Communication and Computation Laboratory (MultiComp Lab). She focuses on computational study of human nonverbal communication. During summer 2011, Derya worked as an intern at AT&T labs in Florham Park, NJ.

Dr. Louis-Philippe Morency is currently research assistant professor at the University of Southern California (USC)and research scientist at USC Institute for Creative Technologies where he leads the Multimodal Communication and Computation Laboratory (MultiComp Lab). He received his Ph.D. from MIT Computer Science and Artificial Intelligence Laboratory in 2006. His main research interest is computational study of human multimodal computation, a multi-disciplinary research topic that overlays the fields of multi-modal interaction, machine learning, computer vision, social psychology and artificial intelligence. He developed Watson, a real-time library for nonverbal behavior recognition and which became the de-facto standard for adding perception to embodied agent interfaces. He received many awards for his work on nonverbal behavior computation including four best papers awards in the last two years (at various IEEE and ACM conferences). He was recently selected by IEEE Intelligent Systems as one of the "Ten to Watch" for the future of AI research.